Meta-analysis of prostate cancer gene expression data identifies a novel discriminatory signature enriched for glycosylating enzymes

Barfeld, S. J., East, P., Zuber, V., & Mills, I. G. (2014). Meta-analysis of prostate cancer gene expression data identifies a novel discriminatory signature enriched for glycosylating enzymes. *BMC Medical Genomics*, *7*, [513]. https://doi.org/10.1186/s12920-014-0074-9

Published in:
BMC Medical Genomics

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

BMC Medical Genomics

BioMed Central
The Open Access Publisher

Meta-analysis of prostate cancer gene expression data identifies a novel discriminatory signature enriched for glycosylating enzymes

BMC Medical Genomics

doi:10.1186/s12920-014-0074-9

Stefan J Barfeld (stefan.barfeld@ncmm.uio.no)
Phil East (phil.east@cancer.org.uk)
Verena Zuber (verena.zuber@ncmm.uio.no)
Ian G Mills (ian.mills@ncmm.uio.no)

Published online: 31 December 2014

ISSN 1755-8794

Article type Research article

Submission date 21 August 2014

Acceptance date 17 December 2014

Article URL http://dx.doi.org/10.1186/s12920-014-0074-9

Meta-analysis of prostate cancer gene expression data identifies a novel discriminatory signature enriched for glycosylating enzymes

Stefan J Barfeld[2*]
[*] Corresponding author
Email: stefan.barfeld@ncmm.uio.no

Phil East[1]
Email: phil.east@cancer.org.uk

Verena Zuber[2]
Email: verena.zuber@ncmm.uio.no

Ian G Mills[2,3,*]
Email: ian.mills@ncmm.uio.no

[1] Bioinformatics & Biostatistics, Cancer Research UK London Research Institute, 44 Lincoln?s Inn Fields, London WC2A 3PX, UK

[2] Prostate Cancer Research Group, Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership University of Oslo and Oslo University Hospital, Oslo, Norway

[3] Department of Cancer Prevention and Urology, Institute of Cancer Research and Oslo University Hospital, Oslo, Norway

Abstract

Background

Tumorigenesis is characterised by changes in transcriptional control. Extensive transcript expression data have been acquired over the last decade and used to classify prostate cancers. Prostate cancer is, however, a heterogeneous multifocal cancer and this poses challenges in identifying robust transcript biomarkers.

Methods

In this study, we have undertaken a meta-analysis of publicly available transcriptomic data spanning datasets and technologies from the last decade and encompassing laser capture microdissected and macrodissected sample sets.

Results

We identified a 33 gene signature that can discriminate between benign tissue controls and localised prostate cancers irrespective of detection platform or dissection status. These genes were significantly overexpressed in localised prostate cancer versus benign tissue in at least

three datasets within the Oncomine Compendium of Expression Array Data. In addition, they were also overexpressed in a recent exon-array dataset as well a prostate cancer RNA-seq dataset generated as part of the The Cancer Genomics Atlas (TCGA) initiative. Biologically, glycosylation was the single enriched process associated with this 33 gene signature, encompassing four glycosylating enzymes. We went on to evaluate the performance of this signature against three individual markers of prostate cancer, v-ets avian erythroblastosis virus E26 oncogene homolog (ERG) expression, prostate specific antigen (PSA) expression and androgen receptor (AR) expression in an additional independent dataset. Our signature had greater discriminatory power than these markers both for localised cancer and metastatic disease relative to benign tissue, or in the case of metastasis, also localised prostate cancer.

Conclusion

In conclusion, robust transcript biomarkers are present within datasets assembled over many years and cohorts and our study provides both examples and a strategy for refining and comparing datasets to obtain additional markers as more data are generated.

Keywords

Transcription, Prostate cancer, Signature

Background

Alterations in transcriptional programmes are often involved in neoplastic transformation and progression and defining these changes will help to understand the underlying biology of the malignancies. Gene Expression Microarray Analysis and more recently high-throughput RNA sequencing (RNA-seq) are commonly used techniques when trying to acquire an unbiased view of the expression levels of large numbers of genes. In order to define more compact and manageable expression modules that might predict risk or prognosis, various approaches have been used across several studies. These include the identification of consensus profiles across multiple datasets [1] and identifying biologically categorised gene sets with enriched representation of deregulated genes [2,3]. Furthermore, smaller expression modules have also been identified using hierarchical clustering methods to generate clusters containing genes with similar expression profiles across glioblastoma samples [3]. The high degree of prostate tissue heterogeneity, however, represents a challenge for transcriptomics since the relative prevalence of each cell type within a given sample determines the overall expression profile. This makes it difficult to compare prostate samples that have very different epithelial and stromal contents. Many studies have compared tumor tissue with benign hyperplastic tissue, or with non-tumoral prostate tissues that were not precisely characterised in terms of location or epithelial representation. Therefore, the outcomes of these analyses were possibly biased because the comparisons included tissues of diverse histological or embryological origins. Various approaches have been used to overcome this issue including *in silico* corrections to compensate for variable epithelial representations in different samples [4], and laser microdissection combined with *in vitro* linear amplification [5]. The laser capture microdissection study of Tomlins *et al.* yielded several informative molecular concepts (multi-gene modules), which provide a rich source of data for further refinement and follow-up as well as distinguishing between stromal and epithelial cancer signatures [5]. It is, however, not clear how detectable those concepts might be in material

extracted from heterogeneous whole tissue sections, an important point given the time and expense associated with laser capture microdissection.

In this study, we have therefore set out with a number of goals. First and foremost amongst these was to determine whether we could identify gene signatures that were statistically significant in datasets generated from both whole tissue sections and laser capture microdissected material. If so, this might indicate that with the right filtering approach, sample heterogeneity might not be a completely confounding challenge to transcriptomic analysis. Secondly, if we were able to identify such signatures, we then wanted to be able to refine them to a point that the signature and any pathway or process enriched within it could be easily validated by other experimental and clinical research groups. Here, we report a concise 33-gene signature with biological enrichment for glycosylation, which discriminates between benign tissue and prostate cancer (PCa) across multiple transcript detection platforms and sample types.

Methods

Description of datasets

Five datasets were downloaded and used in this study.

1. A 19-sample dataset generated by Varambally *et al.,* using the Affymetrix Human Genome U133 Plus 2.0 Array platform. The dataset consisted of 13 macrodissected individual benign prostate, primary and metastatic PCa samples and 6 pooled samples from benign, primary or metastatic PCa tissues. The expression array data were downloaded from GEO under accession number GSE3325 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3325).
2. A 104-sample dataset generated by Tomlins *et al.,* using an in-house cDNA microarray platform (Chinnaiyan Human 20K Hs6). Laser capture microdissection was used to isolate 101 specific cell populations from 44 individuals representing PCa progression in a range of sample categories encompassing 12 stromal and 89 epithelial cell populations. These were subcategorised as EPI_BPH (benign prostatic hyperplasia epithelium), EPI_ADJ_PCA (normal epithelium adjacent to PCa), EPI_ATR (atrophic epithelium ? simple atrophy), EPI_ATR_PIA (atrophic epithelium), PIN (prostatic intraepithelial neoplasia), PCA (prostate carcinoma), MET_HN (Metastatic Prostate Carcinoma - Hormone Naʔve), MET_HR (Metastatic Prostate Carcinoma - Hormone Refractory), STROMA_EPIBPH (BPH Stroma - Epithelial BPH), STROMA_NOR (Normal Stroma - Organ Donor), STROMA_ADJ_PCA (Normal Stroma - Adjacent to prostate cancer). In addition three samples were EPI_NOR (Normal Epithelium - Organ Donor). In our study we maintain this nomenclature in describing the dataset. The expression array data were downloaded from GEO under accession number GSE6099 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6099).
3. A multi-cancer microarray dataset generated by Ramaswamy *et al.,* and consisting of 218 tumour samples, spanning 14 common tumour types, and 90 normal tissue samples and profiled on Affymetrix oligonucleotide microarrays (Hu6800 and Hu35KsubA GeneChips). The 14 tumour types incorporated into this study were breast adenocarcinoma, prostate adenocarcinoma, lung adenocarcinoma, colorectal adenocarcinoma, lymphoma, bladder transitional cell carcinoma, melanoma, uterine

adenocarcinoma, leukemia, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma, pleural mesothelioma and cancers of the central nervous system. The dataset was downloaded from the Broad Institute website (http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=61)

4. A PCa dataset generated by Taylor *et al.,* for 150 tumours, 29 matched normal samples, and 6 cell lines using the Affymetrix Human Exon 1.0 ST array platform. There were 27 metastatic samples amongst the 150 tumours and 35 cases of biochemical relapse (Additional file 1 in Taylor *et al.,*). The expression array data were downloaded from GEO under accession number GSE21034 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21034).
5. High throughput RNA sequencing data were generated by the The Cancer Genomics Atlas (TCGA) consortium for 383 samples, 50 benign samples and 333 primary tumours. 48 of these samples represented advanced disease with Gleason grade ≥8 and 13 cases had undergone progression as characterised by post-operative biochemical recurrence. Data were downloaded from the UCSC Cancer Genome Browser (https://genome-cancer.ucsc.edu/) - TCGA_PRAD_exp_HiSeqV2-2014-05-02.tgz. Associated clinical data were downloaded from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/).
6. A PCa dataset generated by Grasso *et al.,* generated for 28 benign prostate tissue samples, 59 localised PCa and 35 metastatic PCa was generated on two Agilent microarray platforms (whole genome microarray (4x44K ,G4112F) and whole human genome oligo microarray (G4112A). The expression array data were downloaded from GEO under accession number GSE35988 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35988).
7. Additional datasets were interrogated for the expression of individual genes within signatures through the Oncomine Compendium of Expression Array data (www.oncomine.org).

Prostate cancer dependent expression changes

To generate an initial broad progression dependent gene set, we used the prostate progression expression dataset GSE3325 (NCBI GEO database). We quantile normalised probe level intensity values and generated probe set signal estimates using RMA (1,2). We first characterised reporters with a coefficient of variance of less than or equal to 0.05 as uninformative and removed them from further analysis. Reporters having intensities below the 10^{th} quantile (3.91) in more than 75% of the samples were also removed. We identified progression associated expression changes by linear model. Primary tumour versus benign and metastatic versus primary contrasts were run and differential reporters identified using a 0.01 FDR threshold. Reporters were further filtered selecting those with a differential effect size of greater than or equal to 2-fold. This resulted in a progression signature set of 4662 reporters, 3021 genes (121 primary, 2900 metastatic, primary ∩ metastatic = 102). Signatures derived from this primary dataset were subsequently applied to two additional datasets, one prostate dataset generated by Tomlins *et al*. [5] in a laser capture microdissection study (GSE6099) and another generated by Ramaswamy *et al.* [6] and representing multi-tissue primary tumours and metastases (accessible through the Broad Institute data repository: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi).

Identifying correlated gene modules

We clustered our progression gene set using hierarchical clustering with a Ward agglomerative method designed to minimize intra-cluster variance (hclust, Bioconductor) and a 1 - Pearson correlation coefficient dissimilarity measure. We found this method produced a more highly correlated clustering structure when compared to other methods leading to more compact sub-clusters (Additional file 1). We characterised correlated gene modules by cutting the cluster dendrogram at branch lengths ranging from log10(0.05) to log10(3000) giving 39 equal intervals across the log scale. We removed clusters containing less than 3 members from further analysis. We selected modules defined at branch lengths 0f 0.6, 0.8, 1.1, 1.9, 2.5, 4.5, 10.6, 24.7 and 101.6 for further analysis since these gave a broad range of cluster numbers. Since a smaller branch length threshold does not always sub divide a parent module modules can be duplicated at different thresholds. These were removed from further analysis assigning them to the largest branch threshold at which they appears. We assigned Gene Ontology classifications to modules by testing for enrichment at GO nodes using a hyper-geometric distribution and a 0.01 p-value threshold. We carried out this analysis at the gene level by translating chip reporter probeset ids to Entrez gene ids. All reporters from the progression signature with assigned Entrez gene ids were used as background. Analysis was carried out using the GoStats package, Bioconductor.

Phenotype dependent transcript module expression changes

To determine differential regulation of modules within other expression datasets we first identified phenotype dependent expression changes for each sample using an absolute fold change filter of greater than 2. To generate fold changes against which we could filter each gene was scaled to a baseline intensity value. In the case of dataset GSE3325 each signal intensity from the primary tumour samples was scaled to the corresponding median gene signal intensity across the benign tumour samples. Likewise all metastatic samples were scaled to the median across all primary tumour samples. Prior to mapping modules to dataset GSE6099 we background corrected each sample using a normexp method and print tip loess normalised (normalizeWithinArrays(), Bioconductor). We then scaled PIN samples to EPI_ADJ_PCA control samples, PCA samples to PIN, MET_HNF to PCA and MET_HR to PCA. To identify hormone refractory dependent expression changes MET_HR samples were scaled to the median across the non-refractory samples MET_HN. To determine module induction or repression within the scaled samples we tested for enrichment of module genes within the sample associated expression changes using a hypergeometric distribution, <= 0.05 fdr. Mapping was achieved across array platforms using NCBI Entrez gene ids. Modules with an intersection of less than 3 were discarded from the analysis.

Phenotype segregation

To determine if any of the enriched modules were capable of segregating samples on phenotype we built contingency tables across clinical conditions from each of the data sets (Tomlins [5] and Ramaswamy [6]) for induced and repressed modules and tested for sample enrichment using a Fisher?s Exact test. Here we tested each phenotypic group against all others from each data set.

Cluster analysis

To determine the best clustering method and branch length thresholds to apply to our analysis we clustered our 4662 reporter prostate tumour progression signature (Additional file 2: Table S1) using single, average, complete, Ward?s[1] minimum variance method and mcquitty[2] agglomerative hierarchical clustering methods along with a divisive method. The hclust function from Bioconductor[3] was used for the agglomerative techniques and the diana function from the cluster package from Bioconductor was used to run the divisive method. We used the cutree function at branch length thresholds ranging from 0.05 to 2 in increments of 0.05 to derive groups of correlated genes. In the case of the Ward agglomerative method where the branch scales [it is unclear how Ward calculates its branch lengths, need to find out] branch length thresholds ranged from $\log_{10}(0.05)$ to $\log_{10}(3000)$ in increments of $\log_{10}(3000/0.05)/39$. Branch length threshold intervals were chosen to produce a broad range of cluster numbers.

Cluster correlation

To assess the extent to which genes assigned to clusters are correlated we calculated a within-cluster dissimilarity value for each cluster[4]. This is given by

$$W\ C = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2\mathrm{N}_k}\sum_{i=1}^{N}\sum_{j=1}^{N} d\ x_i, x_j$$

where $d\ x_i, x_j$ is the dissimilarity between genes $_i$ and $_j$ across all samples, $i,j = 1,2,\ N$ where N is the total number of cluster members and $k = 1,2,\ ,K$ where K is the total number of clusters. In our case the dissimilarity measure is 1 ? Pearson correlation coefficient between 2 genes across all samples. *W(C)* dissimilarity values across the array of branch length thresholds can be seen plotted against cluster number in Additional file 1: Figure S1. As observed the Ward agglomerative method out performs all other methods producing clusters that are less dissimilar and therefore more highly correlated than those generated from other methods relative to the number of clusters produced. These results provide a justification for using hierarchical clustering with a ward agglomerative method to generate sets of co-regulated genes.

Cluster gene ontology entropy

To quantify the information content of our clusters from a biological perceptive we assigned GO terms to cluster members. This was achieved by mapping GO terms via reporter entrez gene id assignments using the GO.db annotation package from Bioconductor. To quantify the GO information content of a cluster we calculated Shannon Entropy bit values given by:

$$H\ X = -\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N} p\ x_i\ \log_b p x_i$$

where x_i is a cluster associated GO term, $\mathrm{p}(x_i)$ is the probability of choosing x_i from all cluster GO terms , $i = 1,2,\ N$ where N is the total number of unique cluster GO terms, $k = 1,2,\ K$ where K is the total number of clusters and $b = 2$. *H(X)* bit values for different branch length thresholds can be seen plotted against cluster number for the different clustering techniques in Additional file 3: Figure S2. As observed the Ward clustering method produces clusters

with higher GO bit values when compared to other methods. This implies greater uncertainly in the GO term mappings for clusters generated by the ward method thus indicating the production of clusters more GO information rich when compared to other methods. This provides further justification for using hierarchical clustering with a ward agglomerative method to generate sets of co-regulated genes.

Visualization of gene signatures through heatmaps

For visualization, sample groups were averaged using the mean prior to high level mean and variance normalization using the freely available software J-Express 2012 (http://jexpress.bioinfo.no/site/). Subsequently, both sample groups and genes were hierarchically clustered using complete linkage and Euclidian distance using the freely available software Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Heatmaps were produced using Java Tree View (http://jtreeview.sourceforge.net/).

Evaluating gene signature specificity and sensitivity

Testset: Grasso 1 Platform GPL6480

#Title Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
tissue: benign prostate tissue (N)
12
tissue: localised prostate cancer (T)
49
tissue: metastatic castrate resistant prostate cancer (WA)
27.

Training data: Grasso 2 Platform GPL6848

#Title Agilent-012391 Whole Human Genome Oligo Microarray G4112A
tissue: benign prostate tissue (N)
16
tissue: localised prostate cancer (T)
10
tissue: metastatic castrate resistant prostate cancer (WA)
8.

To evaluate the prediction performance 33 gene signature we analysed data from a microarray experiment of Grasso et al as available from GEO (GSE35988) http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35988.

The dataset includes measurements on two different microarray platforms GPL6480 and GPL6848 and includes three different tissue types (benign, localised and metastatic castrate resistant prostate cancer).

We used the samples typed on platform GPL6848 as trainings data to derive the weights of the genes in the multi-gene signature. First, we replaced missing values using the k nearest neighbor algorithm as implemented in the R package impute. The gene PCA3 was not measured on the microarray and the gene CRISP3 was not observed in more than half of these samples. Thus, these two genes were excluded from the signature. We estimated the

weights of each gene using an L2 regularized logit regression model [7] with the R package glmnet.

Then we used the samples typed on platform GPL6480 as test data to evaluate the prediction performance of the gene-signature. Per sample we computed one score as the weighted average over the 31 proposed genes where the weights were defined by the independent training data. Finally, we computed ROC statistics and report the area under the curve (AUC) of the ROC curve (R package ROCR). The AUC indicates the ability of a marker to distinguish between two groups, where a value 0.5 is random and a value of 1 represents a perfect distinction between the groups. Additionally, we were looking at the AUC of specific genes, in particular KLK3, ERG and AR.

Results and discussion

In order to define our starting signatures, we selected a dataset published by Varambally *et al*. about a decade ago and consisting of a small number of whole tissue sections [8]. This constitutes the smallest and oldest dataset used in our meta-analysis. It was, however, extensively validated at both the transcript and protein level in the original study and therefore provides a high-degree of confidence in data quality. We chose to define our starting signatures using this dataset in order to assess how much information could be derived despite the limitations in size and age. Within these data, we firstly identified transcripts that were differentially expressed in localised prostate versus benign tissue or in metastatic disease versus localised cancer using a conventional linear model approach. This approach identified 121 genes differentially expressed in localised primary cancers (primary versus benign, 0.01 FDR) and 2900 genes associated with metastatic status (metastatic versus primary, 0.01 FDR), which were covered by 4662 probes in total (Additional file 2: Table S1). To further refine these gene lists into discrete signatures, we constructed a gene coexpression network using Pearson correlation coefficients and hierarchical clustering using the Ward agglomerative method (**See methods section**).

A number of different correlation or dissimilarity metrics have been employed when constructing co-expression networks. To determine the correlation between the genes we used a Pearson correlation coefficient to construct a dissimilarity matrix across all affected samples in the prostate tumour progression dataset and all genes identified in the preliminary analysis. We then used hierarchical clustering to group the genes. There are a number of available agglomeration methods available each producing their own clustering structure. To determine the best agglomeration method to apply in constructing our expression modules, we clustered our prostate tumour progression signature using single, average, complete, the Ward [9] minimum variance method and the Mcquitty [10] agglomerative hierarchical clustering method along with a divisive method. The performance of these clustering methods by using an algorithm to determine the extent to which genes assigned to clusters are correlated generating a within-cluster dissimilarity value for each cluster (**Methods section Cluster Correlation**). In addition, we assessed the information content in gene ontology terms associated with clusters generated using each method by calculating Shannon Entropy bit values (**Methods section - ?Cluster Gene Ontology Entropy**). Shannon entropy and coefficient of variation are well known in a great many application domains, from theoretical physics to computational chemistry to materials science. They have been applied in bioinformatics as well, most notably in statistical genetics and molecular biology. Shannon entropy is derived from information theory [11]. Most relevant for this study the approach has previously been used as a measure of the robustness of gene regulatory networks [12],to

accelerate feature elimination when classifying microarray expression data [13]. By these measures the Ward clustering method provided both more tightly associated coexpressed gene clusters as well as clusters with higher GO bit values when compared to other methods, indicative of greater information content in the ontologies derived for coexpression clusters generated using the Ward approach than using the other approaches. Additional file 4: Table S2 provides a complete list of coexpressed genes signatures generated used the Ward approach at all branching thresholds.

Four large gene signatures were generated at the least stringent cut-point consisting of 1334 genes referred to as signature 1 (annotated as 101.6.1: Additional file 5: Table S3), 652 genes referred to as signature 2 (annotated as 101.6.2: Additional file 6: Table S4), 836 genes referred to as signature 3 (annotated as 101.6.3: Additional file 7: Table S5) and 357 genes referred to as signature 4 (annotated as 101.6.4: Additional file 8: Table S6). Signatures 1 and 2 contained genes that predominantly discriminated between localised PCa and benign tissue. Using DAVID ontology enrichment search (http://david.abcc.ncifcrf.gov/) to determine whether KEGG pathways were enriched within these signatures, we identified focal adhesions (hsa04510: Focal adhesion, p-value 7.21×10^{-6}) (Table 1) as the most significant pathway for signature 1and complement and coagulation cascades for signature 2 (hsa04610: Complement and coagulation cascades, p-value 0.002) (Table 2). The 38 genes associated with the focal adhesion annotation (p-value 7.21×10^{-6}) in signature 1 are listed in Table 1 and were all significantly downregulated in metastatic samples relative to benign and localised PCa. Half of these genes were laminins (eg. laminin alpha subunit-4 (LAMA4)), integrins (eg. integrin, alpha 1 (ITGA1) and five others), thrombospondins (thrombospondins 1 and 4 (THBS1/4), collagens, actins and myosins which may reflect the remodelling of the extracellular matrix and loss of stroma in particular during the transition to metastasis. The enrichment for complement and coagulation cascades in signature 2 (p value 0.002) included complement (eg. C1R, C1QA, C3) and plasma factors as well as serpin peptidase inhibitor as listed in Table 2 and were also predominantly downregulated in metastatic cases versus benign tissue and localised PCa. Collectively, these pathway enrichments might reflect a combination of extracellular matrix changes and the contribution of infiltrating immune cells and the inflammatory response. However, given that the Varambally dataset consists of whole-tissue sections it is not possible in this meta-analysis to precisely attribute these signatures to a particular biological process.

Table 1 KEGG pathway enrichment analysis for the genes comprising signature 1 (101.6.1)

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa04610:Complement and coagulation cascades	10	1.56	0	C1QA, FGG, A2M, C3, KLKB1, CD46, C1R, SERPING1, C1S, CFD	219	69	5085	3.37	0.33	0.33	2.97
KEGG_PATHWAY	hsa04540:Gap junction	10	1.56	0.01	TJP1, ADCY2, GNAI1, PDGFA, TUBB6, GUCY1A3, GJA1, LPAR1, PRKACB, ITPR2	219	89	5085	2.61	0.88	0.65	14.98
KEGG_PATHWAY	hsa04142:Lysosome	11	1.72	0.03	AGA, HGSNAT, LAMP2, CTSK, GM2A, PSAP, LGMN, CTSB, SCARB2, FUCA1, CLN5	219	117	5085	2.18	0.99	0.77	28.69
KEGG_PATHWAY	hsa04270:Vascular smooth muscle contraction	10	1.56	0.05	PLA2G4A, ADCY2, CALD1, MRVI1, GUCY1A3, PRKCH, PRKACB, PPP1CB, MYLK, ITPR2	219	112	5085	2.07	1	0.87	46.04
KEGG_PATHWAY	hsa04310:Wnt signaling pathway	12	1.88	0.06	CCND1, PRICKLE1, CCND2, BTRC, NFAT5, CAMK2D, TP53, MAPK10, PRKACB, FZD5, FZD4, FZD7	219	151	5085	1.85	1	0.85	51.51
KEGG_PATHWAY	hsa05330:Allograft rejection	5	0.78	0.07	HLA-DRB5, HLA-DPB1, HLA-E, HLA-DOA, HLA-DRA	219	36	5085	3.22	1	0.84	56.25
KEGG_PATHWAY	hsa05416:Viral myocarditis	7	1.1	0.08	CAV1, CCND1, HLA-DRB5, HLA-DPB1, HLA-E, HLA-DOA, HLA-DRA	219	71	5085	2.29	1	0.86	64.57
KEGG_PATHWAY	hsa05332:Graft-versus-host disease	5	0.78	0.08	HLA-DRB5, HLA-DPB1, HLA-E, HLA-DOA, HLA-DRA	219	39	5085	2.98	1	0.82	65.24
KEGG_PATHWAY	hsa04510:Focal adhesion	14	2.19	0.09	CAV1, PDGFA, MAPK10, FLNC, PPP1CB, VCL, CCND1, CCND2, ITGAV, COL6A2, RAP1A, THBS1, PIK3R1, MYLK	219	201	5085	1.62	1	0.81	67.5

Genes comprising signature 1 (Additional file 5: Table S3) were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

Table 2 KEGG pathway enrichment analysis for the genes comprising signature 2 (101.6.2)

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa04110:Cell cycle	36	0.47	9.03E-20	E2F1, E2F2, E2F3, TTK, CHEK1, PTTG1, CCNE2, CCNE1, CDKN2A, MCM7, CDKN2C, CDKN2D, ORC6L, TFDP2, BUB1, CCNA2, STAG1, CDC7, CDC6, RBL1, SKP2, ESPL1, CDC20, MCM2, CDC25C, MCM4, CDC25A, CDC25B, CDKN1C, CCNB1, CCNB2, MAD2L1, PLK1, GSK3B, BUB1B, MAD2L2	225	125	5085	6.51	1.29E-17	1.29E-17	1.07E-16
KEGG_PATHWAY	hsa03030:DNA replication	12	0.16	2.15E-07	RFC5, PRIM1, MCM7, RFC4, POLE2, LIG1, POLA1, POLA2, MCM2, RNASEH2A, MCM4, FEN1	225	36	5085	7.53	3.08E-05	1.54E-05	2.55E-04
KEGG_PATHWAY	hsa04114:Oocyte meiosis	18	0.23	4.82E-06	SGOL1, AURKA, CDC20, ESPL1, PTTG1, CDC25C, CCNE2, CCNB1, CCNE1, CCNB2, MAD2L1, ADCY9, CALML3, PLK1, BUB1, FBXO5, CAMK2B, MAD2L2	225	110	5085	3.7	6.89E-04	2.30E-04	0.01
KEGG_PATHWAY	hsa04914:Progesterone-mediated oocyte maturation	14	0.18	8.26E-05	HSP90AA1, CDC25C, CDC25A, CDC25B, CCNB1, CCNB2, MAD2L1, KRAS, ADCY9, PLK1, BUB1, MAD2L2, PIK3R3, CCNA2	225	86	5085	3.68	0.01	0	0.1
KEGG_PATHWAY	hsa04115:p53 signaling pathway	10	0.13	0	CCNE2, CCNB1, CCNE1, CDKN2A, CCNB2, RRM2, TSC2, CHEK1, PMAIP1, GTSE1	225	68	5085	3.32	0.32	0.07	3.16
KEGG_PATHWAY	hsa05222:Small cell lung cancer	11	0.14	0	E2F1, CCNE2, E2F2, CCNE1, CKS1B, E2F3, PTK2, SKP2, PIAS2, PIK3R3, ITGA2B	225	84	5085	2.96	0.4	0.08	4.12
KEGG_PATHWAY	hsa04360:Axon guidance	14	0.18	0	PLXNA1, EFNB3, PLXNA2, DPYSL5, EPHB1, PTK2, KRAS, UNC5B, PAK2, UNC5A, FYN, GSK3B, SRGAP1, SRGAP2	225	129	5085	2.45	0.45	0.08	4.83
KEGG_PATHWAY	hsa00240:Pyrimidine metabolism	11	0.14	0.01	PRIM1, TYMS, POLR3K, POLE2, RRM2, RRM1, DCK, POLA1, POLA2, NME7, TK1	225	95	5085	2.62	0.71	0.14	9.63
KEGG_PATHWAY	hsa05219:Bladder cancer	7	0.09	0.01	E2F1, RPS6KA5, E2F2, E2F3, CDKN2A, KRAS, PGF	225	42	5085	3.77	0.74	0.14	10.67
KEGG_PATHWAY	hsa05215:Prostate cancer	10	0.13	0.02	E2F1, CCNE2, E2F2, CCNE1, E2F3, HSP90AA1, KRAS, GSK3B, PIK3R3, CTNNB1	225	89	5085	2.54	0.9	0.2	17.14
KEGG_PATHWAY	hsa00230:Purine metabolism	14	0.18	0.02	POLR3K, POLA1, DCK, POLA2, HPRT1, GMPS, NME7, GART, PRIM1, ADCY9, POLE2, RRM2, PKLR, RRM1	225	153	5085	2.07	0.91	0.2	18.08
KEGG_PATHWAY	hsa03410:Base excision repair	6	0.08	0.02	POLE2, UNG, LIG1, MBD4, NTHL1, FEN1	225	35	5085	3.87	0.92	0.19	18.86
KEGG_PATHWAY	hsa05214:Glioma	8	0.1	0.02	E2F1, E2F2, E2F3, CDKN2A, KRAS, CALML3, CAMK2B, PIK3R3	225	63	5085	2.87	0.94	0.2	21.16

KEGG_PATHWAY	hsa05200:Pathways in cancer	23	0.3	0.03	E2F1, E2F2, FZD8, CKS1B, MSH6, E2F3, HSP90AA1, PGF, FGF9, SKP2, BIRC5, FZD2, CTNNB1, CTNNA2, CCNE2, CCNE1, PTK2, CDKN2A, KRAS, GSK3B, PIAS2, PIK3R3, ITGA2B	225	328	5085	1.58	0.99	0.27	30.48
KEGG_PATHWAY	hsa00670:One carbon pool by folate	4	0.05	0.03	TYMS, MTHFD2, SHMT2, GART	225	16	5085	5.65	0.99	0.26	31.03
KEGG_PATHWAY	hsa04916:Melanogenesis	9	0.12	0.07	FZD8, KRAS, ADCY9, CALML3, GSK3B, GNAS, CAMK2B, FZD2, CTNNB1	225	99	5085	2.05	1	0.47	57.19
KEGG_PATHWAY	hsa05210:Colorectal cancer	8	0.1	0.08	FZD8, MSH6, KRAS, GSK3B, BIRC5, FZD2, PIK3R3, CTNNB1	225	84	5085	2.15	1	0.48	60.57
KEGG_PATHWAY	hsa03430:Mismatch repair	4	0.05	0.08	RFC5, MSH6, RFC4, LIG1	225	23	5085	3.93	1	0.48	61.82
KEGG_PATHWAY	hsa05223:Non-small cell lung cancer	6	0.08	0.09	E2F1, E2F2, E2F3, CDKN2A, KRAS, PIK3R3	225	54	5085	2.51	1	0.5	66.23
KEGG_PATHWAY	hsa05218:Melanoma	7	0.09	0.09	E2F1, E2F2, E2F3, CDKN2A, KRAS, FGF9, PIK3R3	225	71	5085	2.23	1	0.5	67.79
KEGG_PATHWAY	hsa05212:Pancreatic cancer	7	0.09	0.1	E2F1, E2F2, E2F3, CDKN2A, KRAS, PGF, PIK3R3	225	72	5085	2.2	1	0.5	69.77

Genes comprising signature 2 (Additional file 6: Table S4) were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

By contrast, signatures 3 and 4 contained genes that predominantly discriminated between metastatic cases and benign tissue samples. The dominant pathway for signature 3 was cell cycle regulation (hsa04110: Cell cycle, p-value $9x10^{-20}$) and the enrichment arose from the overexpression of a total of 36 genes linked to this process in the metastatic cases versus benign tissue. The genes are listed in Table 3 and included E2F transcription factors, DNA replication licensing factors, cyclin-dependent kinase inhibitors, cell division cycle genes and components of the mitotic spindle checkpoint control apparatus. Many of these overexpressed genes also constitute a prognostic cell cycle progression gene signature, which has been validated at the transcript level in biopsy samples [14]). For signature 4, steroid biosynthesis was the most enriched pathway (hsa00100: Steroid biosynthesis, p-value 0.03 ? squalene epoxidase (SQLE), farnesyl-diphosphate farnesyltransferase 1 (FDFT1), sterol-C4-methyl oxidase-like gene (SC4MOL). In this case the enrichment was due to the differential expression of three genes that are functionally tightly linked in some cases on consecutive steps in the cholesterol biosynthesis pathway. FDFT1 was overexpressed, SC4MOL was downregulated and SQLE showed a switch in expression in which one probe on the array was repressed and another was overexpressed (Table 4). Downregulation occured predominantly in localized PCa relative to benign tissue and expression seemed higher in metastatic cases than localized prostate cancers. . FDFT1 overexpression, and increases in the expression of one probe for SQLE, were most significant in the metastatic cases compared to benign tissue and localised disease. These are enzymes associated with cholesterol biosynthesis in particular and collectively catalyse 3 out of 4 consecutive reactions in the conversion of farnesyl pyrophosphate to lathosterol via squalene. FDFT1 catalyses the production of squalene from farnesyl pyrophosphate, SQLE catalyses the conversion of squalene to 2,3-epoxysqualene and SC4MOL catalyses the conversion of lanosterin to lathosterol. The two metabolites consecutively further downstream in the pathway are dehydrocholesterol and cholesterol. FDFT1 overexpression has previously been associated with aggressive PCa [15]). This is particularly intriguing since metastatic PCa is characterised by increases in the proliferative index of tumours [16] and the ability to produce autocrine steroid hormones from cholesterol in order to maintain androgen receptor activity [17]. Consequently the observation of increased levels of these enzymes in metastatic cases may hypothetically imply enhanced cholesterol biosynthesis to sustain its use for steroid hormone biogenesis by the tumours.

Table 3 KEGG pathway enrichment analysis for the genes comprising signature 3 (101.6.3)

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa04110:Cell cycle	36	0.47	9.03E-20	E2F1, E2F2, E2F3, TTK, CHEK1, PTTG1, CCNE2, CCNE1, CDKN2A, MCM7, CDKN2C, CDKN2D, ORC6L, TFDP2, BUB1, CCNA2, STAG1, CDC7, CDC6, RBL1, SKP2, ESPL1, CDC20, MCM2, CDC25C, MCM4, CDC25A, CDC25B, CDKN1C, CCNB1, CCNB2, MAD2L1, PLK1, GSK3B, BUB1B, MAD2L2	225	125	5085	6.51	1.29E-17	1.29E-17	1.07E-16
KEGG_PATHWAY	hsa03030:DNA replication	12	0.16	2.15E-07	RFC5, PRIM1, MCM7, RFC4, POLE2, LIG1, POLA1, POLA2, MCM2, RNASEH2A, MCM4, FEN1	225	36	5085	7.53	3.08E-05	1.54E-05	2.55E-04
KEGG_PATHWAY	hsa04114:Oocyte meiosis	18	0.23	4.82E-06	SGOL1, AURKA, CDC20, ESPL1, PTTG1, CDC25C, CCNE2, CCNB1, CCNE1, CCNB2, MAD2L1, ADCY9, CALML3, PLK1, BUB1, FBXO5, CAMK2B, MAD2L2	225	110	5085	3.7	6.89E-04	2.30E-04	0.01
KEGG_PATHWAY	hsa04914:Progesterone-mediated oocyte maturation	14	0.18	8.26E-05	HSP90AA1, CDC25C, CDC25A, CDC25B, CCNB1, CCNB2, MAD2L1, KRAS, ADCY9, PLK1, BUB1, MAD2L2, PIK3R3, CCNA2	225	86	5085	3.68	0.01	0	0.1
KEGG_PATHWAY	hsa04115:p53 signaling pathway	10	0.13	0	CCNE2, CCNB1, CCNE1, CDKN2A, CCNB2, RRM2, TSC2, CHEK1, PMAIP1, GTSE1	225	68	5085	3.32	0.32	0.07	3.16
KEGG_PATHWAY	hsa05222:Small cell lung cancer	11	0.14	0	E2F1, CCNE2, E2F2, CCNE1, CKS1B, E2F3, PTK2, SKP2, PIAS2, PIK3R3, ITGA2B	225	84	5085	2.96	0.4	0.08	4.12
KEGG_PATHWAY	hsa04360:Axon guidance	14	0.18	0	PLXNA1, EFNB3, PLXNA2, DPYSL5, EPHB1, PTK2, KRAS, UNC5B, PAK2, UNC5A, FYN, GSK3B, SRGAP1, SRGAP2	225	129	5085	2.45	0.45	0.08	4.83
KEGG_PATHWAY	hsa00240:Pyrimidine metabolism	11	0.14	0.01	PRIM1, TYMS, POLR3K, POLE2, RRM2, RRM1, DCK, POLA1, POLA2, NME7, TK1	225	95	5085	2.62	0.71	0.14	9.63
KEGG_PATHWAY	hsa05219:Bladder cancer	7	0.09	0.01	E2F1, RPS6KA5, E2F2, E2F3, CDKN2A, KRAS, PGF	225	42	5085	3.77	0.74	0.14	10.67

KEGG_PATHWAY	hsa05215:Prostate cancer	10	0.13	0.02	E2F1, CCNE2, E2F2, CCNE1, E2F3, HSP90AA1, KRAS, GSK3B, PIK3R3, CTNNB1	225	89	5085	2.54	0.9	0.2	17.14
KEGG_PATHWAY	hsa00230:Purine metabolism	14	0.18	0.02	POLR3K, POLA1, DCK, POLA2, HPRT1, GMPS, NME7, GART, PRIM1, ADCY9, POLE2, RRM2, PKLR, RRM1	225	153	5085	2.07	0.91	0.2	18.08
KEGG_PATHWAY	hsa03410:Base excision repair	6	0.08	0.02	POLE2, UNG, LIG1, MBD4, NTHL1, FEN1	225	35	5085	3.87	0.92	0.19	18.86
KEGG_PATHWAY	hsa05214:Glioma	8	0.1	0.02	E2F1, E2F2, E2F3, CDKN2A, KRAS, CALML3, CAMK2B, PIK3R3	225	63	5085	2.87	0.94	0.2	21.16
KEGG_PATHWAY	hsa05200:Pathways in cancer	23	0.3	0.03	E2F1, E2F2, FZD8, CKS1B, MSH6, E2F3, HSP90AA1, PGF, FGF9, SKP2, BIRC5, FZD2, CTNNB1, CTNNA2, CCNE2, CCNE1, PTK2, CDKN2A, KRAS, GSK3B, PIAS2, PIK3R3, ITGA2B	225	328	5085	1.58	0.99	0.27	30.48
KEGG_PATHWAY	hsa00670:One carbon pool by folate	4	0.05	0.03	TYMS, MTHFD2, SHMT2, GART	225	16	5085	5.65	0.99	0.26	31.03
KEGG_PATHWAY	hsa04916:Melanogenesis	9	0.12	0.07	FZD8, KRAS, ADCY9, CALML3, GSK3B, GNAS, CAMK2B, FZD2, CTNNB1	225	99	5085	2.05	1	0.47	57.19
KEGG_PATHWAY	hsa05210:Colorectal cancer	8	0.1	0.08	FZD8, MSH6, KRAS, GSK3B, BIRC5, FZD2, PIK3R3, CTNNB1	225	84	5085	2.15	1	0.48	60.57
KEGG_PATHWAY	hsa03430:Mismatch repair	4	0.05	0.08	RFC5, MSH6, RFC4, LIG1	225	23	5085	3.93	1	0.48	61.82
KEGG_PATHWAY	hsa05223:Non-small cell lung cancer	6	0.08	0.09	E2F1, E2F2, E2F3, CDKN2A, KRAS, PIK3R3	225	54	5085	2.51	1	0.5	66.23
KEGG_PATHWAY	hsa05218:Melanoma	7	0.09	0.09	E2F1, E2F2, E2F3, CDKN2A, KRAS, FGF9, PIK3R3	225	71	5085	2.23	1	0.5	67.79
KEGG_PATHWAY	hsa05212:Pancreatic cancer	7	0.09	0.1	E2F1, E2F2, E2F3, CDKN2A, KRAS, PGF, PIK3R3	225	72	5085	2.2	1	0.5	69.77

Genes comprising signature 3 (Additional file 7: Table S5) were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

Table 4 KEGG pathway enrichment analysis for the genes comprising signature 3 (101.6.4)

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa00100:Steroid biosynthesis	3	0.1	0.03	SQLE, FDFT1, SC4MOL	86	17	5085	10.43	0.97	0.97	30.81
KEGG_PATHWAY	hsa05200:Pathways in cancer	11	0.38	0.05	LAMA1, HRAS, PTK2, SOS1, CBL, VEGFA, PPARG, RALA, LEF1, MDM2, LAMB1	86	328	5085	1.98	0.99	0.93	41.08
KEGG_PATHWAY	hsa04510:Focal adhesion	8	0.27	0.05	LAMA1, HRAS, PTK2, FLT1, DIAPH1, SOS1, VEGFA, LAMB1	86	201	5085	2.35	1	0.85	44.04
KEGG_PATHWAY	hsa00330:Arginine and proline metabolism	4	0.14	0.06	ARG1, P4HA2, P4HA1, CPS1	86	53	5085	4.46	1	0.82	49.31
KEGG_PATHWAY	hsa05216:Thyroid cancer	3	0.1	0.08	HRAS, PPARG, LEF1	86	29	5085	6.12	1	0.86	63

Genes comprising signature 4 (Additional file 8: Table S6) were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

Discrimination between cancer and benign control tissue and also between metastatic disease and other clinical cases represents an important goal of biomarker research. Thus, we used these gene signatures to classify clinical samples in prostate cancer samples and metastatic tissue samples in two additional datasets. One consisted of prostate cancer samples isolated by laser capture microdissection generated by Tomlins *et al.* [5] and the other contained expression array data from primary and metastatic tumours from multiple tissue sites generated by Ramaswamy *et al.* [6].

The Tomlins dataset consisted of various refined subgroups based on isolation of cell sub-populations including stromal fractions, epithelial fractions, localised prostate cancer and hormone-na?ve and refractory metastatic disease. The Ramaswamy *et al.* dataset consisted of cancers from 14 organ sites with paired normal samples as well as normal tissues. In each dataset, we asked whether our signatures and sub-signatures could discriminate between the sample groups. To determine this, we first assessed the mean fold-change in the expression of each gene signature in each sample group in both datasets (Additional file 9: Table S7). We then performed a Fischer?s Exact test to identify signatures that were capable of discriminating between localised prostate cancer, metastatic prostate cancer and the other sample groups defined in all three published studies ? Varambally *et al.,* Tomlins e*t al.,* and Ramaswamy *et al.* (refer to Materials and Methods for more detail on subgroups/sample types) (Additional file 10: Table S8). Gene ontologies were assigned to these statistically significant gene clusters and the clustering is represented in a heatmap for the classifying modules combining both the Tomlins and Ramaswamy sample sets (Figure 1 and Additional files 10 and 11: Tables S8 and S9 for gene ontology annotations). The smallest gene signature (dist.0.6.34) capable of subclustering localised prostate cancer from other samples in all three datasets consisted of 71 genes (Table 5). This small signature was a sub-component of the original signature 1 (101.6.1). The most significantly enriched biological process associated with these genes was vascular smooth muscle contraction (hsa04270: Vascular smooth muscle contraction, p-value $2x10^{-3}$) (Table 6). The four genes within this signature that were individually most significantly overexpressed in localised prostate cancers compared to benign tissues and metastatic cases were an oncogenic transcription factor, v-myc avian myelocytomatosis viral oncogene homolog (MYC), a proteoglycan capable of sequestering transforming growth factor beta called fibromodulin (FMOD), a mitochondrial enzyme associated with fatty acid metabolism called glycine N-acyltransferase-like protein 1 (GLYATL1) and an extraneuronal monoamine transporter called solute carrier family 22 member 3 (SLC22A3). MYC has been shown to be overexpressed in prostate cancer [18] and to drive tumourigenesis in a transgenic model of the disease [19]. Fibromodulin has not been widely studied in cancer and has not been implicated in prostate cancer. It is, however, known to be significantly overexpressed in chronic lymphocytic leukemia (CLL) versus normal B lymphocytes [20] and associated with a resistance signature to DNA damage-induced apoptosis in CLL [21]. Furthermore, the expression of fibromodulin is known to be induced in leiomyoma in response to TGF-beta through Smad and MAP kinase signalling [22]. GLYATL1 has not been associated with cancers. SLC22A3 has been reported to be overexpressed in localised prostate cancer at the transcript level when compared to benign tissue [5].

Figure 1 Gene signatures capable of discriminating between prostate cancer subgroups and classify metastatic disease. Gene signatures generated using the Varambally dataset and found to be significant discriminators of metastatic disease and primary/localised cancers (Additional file 10: Table S8) when applied to the Tomlins and Rawaswamy datasets were used to cluster samples in these datasets in a heatmap. The gene signatures represented are those capable of characterising samples from at least one progression stage (Fischer?s exact < = 0.05). Gene signatures are rows and samples are columns. The colour coded bar at the base of the heatmap indicates the clinical grouping for each sample as also defined in the key. Metastatic hormone refractory, metastatic hormone na?ve and hormone refractory vs. na?ve represent prostate cancer cases from the Tomlins dataset, as do PIN (prostatic intraepithelial neoplasia) and primary carcinoma. The other categories (metastatic and primary) are samples from the Rawaswamy dataset and are metastatic and primary cancers from multiple organ sites, not simply the prostate gland. The blue bar graph on the right-hand side of the heatmap depicts the number of genes in each signature which are differentially expressed and contribute to the sample clustering in this analysis. For signature 1 (dist 101.6.1 and Additional file 5: Table S3) this is 1748 genes in total as highlighted and other bars are numbers of genes relative to this. The colour scale represents the mean log2 fold change for differential gene signatures (> = abs log2(2)). Red indicates module induction, green repression. Gene signatures significant in both directions are indicated in yellow generated by the addition of the corresponding red and green shades. Using the mean module log2 fold change we clustered the samples and modules using hierarchical clustering with euclidean distance as a measure of dissimilarity. Data points that contained both induced and repressed values have been excluded from the clustering.

Table 5 A 71-gene signature capable of subclustering localised prostate cancer cases across multiple datasets

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa04270:Vascular smooth muscle contraction	5	6.94	0	ACTG2, MYH11, KCNMB1, MYLK, MYL9	26	112	5085	8.73	0.12	0.12	1.99
KEGG_PATHWAY	hsa05414:Dilated cardiomyopathy	4	5.56	0.01	DES, PLN, IGF1, TPM2	26	92	5085	8.5	0.47	0.27	9.6
KEGG_PATHWAY	hsa04960:Aldosterone-regulated sodium reabsorption	3	4.17	0.02	IGF1, ATP1A2, IRS1	26	41	5085	14.31	0.66	0.31	15.92
KEGG_PATHWAY	hsa04310:Wnt signaling pathway	4	5.56	0.04	SFRP1, CAMK2G, PRICKLE2, MYC	26	151	5085	5.18	0.91	0.45	31.53
KEGG_PATHWAY	hsa05410:Hypertrophic cardiomyopathy (HCM)	3	4.17	0.06	DES, IGF1, TPM2	26	85	5085	6.9	0.99	0.57	49.28

A 71-gene signature representing a subset of genes from signature 1 (101.6.1). Columns 1 and 2 represent Affymetrix probe identifiers and gene symbols, respectively. Columns 3-5 represent probe signals in benign, localised prostate cancer and metastatic samples, respectively. Column 6 (?primary.reporters) represents the differential probe signal between benign and localised prostate cancer cases with the associated p-values in column 7. Column 8 (?metastatic.reporters?) represents the differential probe signal between metastatic and localised prostate cases with the associated p-values in column 9. Subsequent columns headed dist.0.6 through to dist.101.6 represent the gene signature codes for coexpressed genes at diminishing significance thresholds proceeding left-to-right across the table. All 71 genes form part of the same signature at all stringency thresholds and consequently the numbering in these columns for all genes is the same. Subsequent columns provide the gene location and gene name including information on chromosome number, cytogenetic band identifier and unigene accession codes. The same table layout is used in Additional files 2, 4, 5, 6, 7 and 8: Tables S1-S6.

Table 6 71-gene signature capable of subclustering localised prostate cancer cases across multiple datasets

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	hsa00512:O-Glycan biosynthesis	3	0.34	0.01	GALNTL4, GCNT1, ST6GALNAC1	26	30	5085	19.56	0.43	0.43	8.95
KEGG_PATHWAY	hsa04610:Complement and coagulation cascades	3	0.34	0.04	C4A, C4B, SERPINA1	26	69	5085	8.5	0.94	0.75	36.76
KEGG_PATHWAY	hsa05322:Systemic lupus erythematosus	3	0.34	0.08	C4A, C4B, HLA-DMB	26	99	5085	5.93	1	0.83	58.75

Genes were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

The other genes within this coexpression signature were downregulated in prostate cancers versus benign tissue and the majority were myosins, such as myosin, heavy polypeptide 11, smooth muscle (MYH11), myocardin (MYOCD), and myosin, light chain 9, regulatory (MYL9)thus accounting for the pathway enrichment for vascular smooth muscle contraction. As prostate cancer progresses to more advanced stages there is a depletion of stromal cells from the tissue and this perhaps explains the dominant contribution from downregulated muscle-associated genes to the signature and also other features of pathway enrichments particularly of the focal adhesion classification [23]. In order to determine whether our signature was consistent across more recent datasets, we downloaded an exon-array dataset generated by Taylor *et al.,* and also The Cancer Genome Atlas (TCGA) data recently generated using high-throughput transcript sequencing of prostate cancers [24] (data generated by the data generated by the TCGA Research Network: http://cancergenome.nih.gov/). MYC and GLYATL1 remain significantly overexpressed features (>1.3 fold) within these signatures in both datasets (Figure 2) with the vast majority of other gene transcripts downregulated including those enriched in the KEGG pathway analysis for vascular smooth muscle contraction.

Figure 2 Differential expression of a 71-gene signature classifier in a prostate cancer exon-array dataset (Taylor *et al.*) and the TCGA RNA-seq dataset for prostate cancer (TCGA-PRAD). The expression values of the 71-gene signature (dist.0.6.34) capable of subclustering localised prostate cancer from other samples in all three interrogated datasets are shown in two independent datasets, **A.** a prostate cancer exon-array dataset (Taylor *et al.*) and **B.** TCGA RNA-seq dataset for prostate cancer (TCGA-PRAD) were used. Values were log2 normalized and the mean of the sample groups (PRIMARY TUMOUR/SOLID TISSUE NORMAL) is shown.

Whilst our 71 gene signature mainly contains differentially expressed genes that are downregulated in cancers versus benign tissues, most prostate cancer biomarkers that are currently under evaluation are overexpressed transcripts and proteins in the disease state. Consequently, we next sought to evaluate genes that were overexpressed in localised prostate cancers in signatures 1-4 more thoroughly in other datasets. There were 97 annotated gene transcripts in total overexpressed (Table 7). We had previously performed pathway analyses on signatures 1-4 which included both up- and downregulated genes (Tables 1, 2, 3 and 4). We now repeated this solely for the 97 overexpressed genes and this yielded pathway enrichment for O-glycan biosynthesis (hsa00512: O-glycan biosynthesis, p-value 0.009) as the most significant KEGG enrichment (Table 8).

Table 7 Overexpressed genes in localised prostate cancer versus benign tissue within the entire set of differentially expressed genes in the Varambally dataset (GSE3325)

Gene/signature	AUC benign-local	AUC benign-metastatic	AUC localized-metastatic
KLK3	0.5204082	0.9104938	0.8707483
ERG	0.812616	0.9326599	0.6099773
AR	0.6581633	0.8395062	0.8435374
31 Gene signature	0.994898	0.9938272	0.957672
Derived from the Grasso Data			

Genes overexpressed in localised prostate cancer are ranked according to the degree of overexpression from highest to lowest within dataset GSE3325. Genes highlighted in **bold** are also overexpressed in at least three other independent prostate cancer expression array datasets hosted by the Oncomine compendium (http://www.oncomine.org) within the Top 1% of overexpressed gene transcripts in all cases.

Table 8 KEGG pathway enrichment analysis for the entire set of overexpressed genes in localised prostate cancer versus benign tissue in the Varambally dataset (GSE3325)

	id	agilent_probe	gene	auc.benign.vs.local	auc.benign.vs.met	auc.local.vs.met
1	10308	A_23_P334538	KLK3	0.668367346938776	0.512345679012346	0.583522297808012
2	18945	A_24_P108401	KLK3	0.517006802721089	0.651234567901235	0.662131519274376
3	24972	A_24_P344510	KLK3	0.520408163265306	0.910493827160494	0.870748299319728
4	9045	A_23_P301414	ERG	0.780487804878049	0.832167832167832	0.578799249530957
5	15382	A_23_P57323	ERG	0.790816326530612	0.759259259259259	0.507180650037793
6	31809	A_24_P922378	ERG	0.812615955473098	0.932659932659933	0.609977324263039
7	1039	A_23_P113111	AR	0.658163265306122	0.839506172839506	0.843537414965986
8	18899	A_24_P106297	AMACR	0.952380952380952	0.771604938271605	0.606198034769463
9	5760	A_23_P1833	B3GAT1	0.741496598639456	0.771604938271605	0.835222978080121
10	31242	A_24_P914625	BEND4	0.82312925170068	0.583333333333333	0.689342403628118
11	2222	A_23_P128304	BICD1	0.894557823129252	0.830246913580247	0.55026455026455
12	31375	A_24_P916586	BICD1	0.852040816326531	0.62962962962963	0.623582766439909
13	38999	A_32_P472968	BICD1	0.906462585034014	0.632716049382716	0.866969009826153
14	31242	A_24_P914625	BEND4	0.82312925170068	0.583333333333333	0.689342403628118
15	14293	A_23_P45786	COL9A2	0.714285714285714	0.521604938271605	0.574452003023432
16	13511	A_23_P419760	CRISP3	1	0.8	0.6
17	12893	A_23_P403466	DLX1	0.743197278911565	0.935185185185185	0.808767951625095
18	34905	A_32_P142818	DLX1	0.985157699443414	0.956228956228956	0.647770219198791
19	18202	A_23_P93058	DNAH5	0.836734693877551	0.62037037037037	0.616780045351474
20	19295	A_24_P120462	DNAH5	0.706439393939394	0.516528925619835	0.650568181818182
21	24667	A_24_P333461	DNAH5	0.74113475177305	0.59	0.542978723404255
22	26067	A_24_P388786	DNAH5	0.870748299319728	0.503086419753086	0.714285714285714
23	7096	A_23_P213424	ENC1	0.698979591836735	0.709876543209877	0.529100529100529
24	29137	A_24_P69095	ENC1	0.819727891156463	0.805555555555556	0.932728647014361
25	18033	A_23_P91081	EPCAM	0.819727891156463	0.709876543209877	0.525321239606954
26	9045	A_23_P301414	ERG	0.780487804878049	0.832167832167832	0.578799249530957
27	15382	A_23_P57323	ERG	0.790816326530612	0.759259259259259	0.507180650037793
28	31809	A_24_P922378	ERG	0.812615955473098	0.932659932659933	0.609977324263039
29	13214	A_23_P412214	RAP1GAP2	0.654761904761905	0.521604938271605	0.643990929705216
30	31091	A_24_P911927	RAP1GAP2	0.62037037037037	0.753623188405797	0.678743961352657
31	18139	A_23_P9232	GCNT1	0.945578231292517	0.604938271604938	0.6432350718065

32	5181	A_23_P16523	GDF15	0.772108843537415	0.549382716049383	0.613756613756614
33	7901	A_23_P250444	GJB1	0.86734693877551	0.62037037037037	0.628873771730915
34	26494	A_24_P404840	GJB1	0.789115646258503	0.503086419753086	0.737717309145881
35	11669	A_23_P370666	GLYATL1	0.857142857142857	0.558641975308642	0.806500377928949
36	7727	A_23_P2344	HOXC6	0.883333333333333	0.863636363636364	0.533333333333333
37	147	A_23_P101806	HPN	0.935374149659864	0.817901234567901	0.595616024187453
38	13011	A_23_P406782	HPN	0.88265306122449	0.722222222222222	0.53817082388511
39	6231	A_23_P203933	ITPR2	0.683673469387755	0.777777777777778	0.869992441421013
40	9378	A_23_P310	MARCKSL1	0.892857142857143	0.953703703703704	0.699168556311413
41	2744	A_23_P13442	MICAL2	0.502777777777778	0.638095238095238	0.642857142857143
42	7851	A_23_P24843	MICAL2	0.821428571428571	0.820987654320988	0.522297808012094
43	7314	A_23_P215956	MYC	0.909863945578231	0.595679012345679	0.689342403628118
44	20716	A_24_P178011	MYC	0.530612244897959	0.87037037037037	0.835222978080121
45	25948	A_24_P38363	MYC	0.772727272727273	0.975	0.656060606060606
46	3126	A_23_P139327	OR51E2	0.903061224489796	0.645061728395062	0.816326530612245
47	22188	A_24_P235756	OR51E2	0.840136054421769	0.737654320987654	0.862433862433862
48	8971	A_23_P29816	PLA1A	0.758503401360544	0.734567901234568	0.808767951625095
49	23649	A_24_P294408	PLA1A	0.811224489795918	0.799382716049383	0.843537414965986
50	15416	A_23_P5778	RAB17	0.928571428571429	0.925925925925926	0.606953892668178
51	7973	A_23_P251387	REPS2	0.884353741496599	0.509259259259259	0.764928193499622
52	33496	A_32_P100109	REPS2	0.848639455782313	0.891975308641975	0.925925925925926
53	6876	A_23_P211110	SIM2	0.797619047619048	0.811728395061728	0.562358276643991
54	9061	A_23_P301886	SIM2	0.807291666666667	0.788461538461538	0.50400641025641
55	15035	A_23_P52939	SLC43A1	0.857142857142857	0.657407407407407	0.53514739229025
56	22855	A_24_P260443	THBS4	0.860544217687075	0.808641975308642	0.963718820861678
57	1849	A_23_P123503	TRIB1	0.770408163265306	0.654320987654321	0.569916855631141
58	22636	A_24_P252497	TRIB1	0.698979591836735	0.524691358024691	0.690854119425548
59	4805	A_23_P160460	UAP1	0.841836734693878	0.657407407407407	0.854875283446712

Genes were uploaded into the DAVID gene ontology search engine (http://david.abcc.ncifcrf.gov/). KEGG pathway enrichment was generated and the table represents the output file ranked based on significance and annotated by column header.

To further refine this gene set, we then interrogated the Oncomine compendium of expression array data to determine which of these 97 genes are significantly overexpressed in at least three additional independent prostate cancer datasets when a Top 1% overexpression threshold was applied together with a p-value threshold of $1x10^{-4}$ [25]. Thirty three annotated genes, around one-third of the 97-gene set fulfilled these criteria. This included 3 of the 4 overexpressed genes (MYC, GLYATL1 and SLC22A3) in the 71 gene signature subgrouping prostate cancers in the Varambally, Tomlins and Ramaswamy studies (highlighted in bold in Table 7). This 33 gene set also included four of the five glycosylating enzymes (UDP N-acetylglucosamine pyrophosphorylase 1 (UAP1), glucosaminyl (N-acetyl) transferase 1, core 2 (GCNT1), beta-1,3-glucuronyltransferase 1 (B3GAT1) and RAP1 GTPase activating protein 2 (RAP1GAP2/GARNL4)) contributing to the ontology enrichment for glycan biosynthesis in the larger 97 gene set. Notably, others and we have recently reported that UAP1 and GCNT1 are overexpressed in prostate cancer tissue using immunohistochemistry. In addition, an aminosugar conjugate, O-linked N-acetylglucosamine (O-GlcNAc), is also significantly elevated in prostate cancer [26,27]. Furthermore, the UAP1 transcript has also been reported to be detectable in urine and plasma samples as a component of a multi-gene signature [28]. Additionally, UDP sugar conjugates have been identified as elevated in prostate cancers through metabolomics and O-linked N-acetylglucosamine is an overexpressed prostate cancer tissue biomarker, which can be conjugated to a variety of proteins to affect their stability and activity including c-Myc [29,30]. Consequently, the presence of these genes encoding glycosylating enzymes within this signature has been partly validated in tissue at the proteins level and suggests that more systematic profiling of glycoproteins may reveal new biomarkers.

Biologically, it is interesting to consider what might contribute to the increased expression of these genes in prostate cancers. Prostate cancer is driven by the dysregulated expression and activity of a number of transcription factors. The most notable example is the androgen receptor but others are overexpressed through chromosomal rearrangements and gene fusions as well as copy number variation as prostate cancer develops and progresses. This in turn has a significant impact on the expression of gene targets for these transcription factors and makes it plausible that a proportion of overexpressed genes reflect changes in transcription factor expression and activity. In this context, it is noteworthy that a total of five transcription factors were present in this group of 33 annotated genes ((single-minded family bHLH transcription factor 2(SIM2), MYC, distal-less homeobox 1 (DLX1), homeobox C6 (HOXC6) and v-ets avian erythroblastosis virus E26 oncogene homolog (ERG)).

c-Myc is a well-established oncogenic transcription factor, which is overexpressed through chromosomal amplification on 8q24 but also through post-translational events, which may include glycosylation of the N-terminal transactivation and concomitant antagonism of proteasomal degradation [31,32] [18]. ERG is part of a highly prevalent gene fusion affecting chromosome 21 and driven by the activity of the AR [33]. It is overexpressed in around 50% of prostate cancers through a chromosomal rearrangement, which fuses it to the upstream androgen receptor-dependent regulatory element controlling TMPRSS2 expression. SIM2 overexpression is associated with changes in transcriptional control affecting other loci on chromosome 21 [34,35]. Target genes for MYC and ERG have been extensively explored in clinical and cell-line datasets using expression array profiling with targeted knockdown and overexpression in prostate cancer cells. These approaches have linked MYC to processes including ribosome biogenesis and splicing and ERG to cell motility and migration, respectively [36-38]. Whilst the 33 genes did not include significant number of established MYC target genes, several reported ERG target were present including B3GAT1,

phospholipase A1 member A (PLA1A) and collagen, type IX, alpha 2 (COL9A2) [39]. In addition, there were a number of direct AR targets including UAP1 and GCNT [30,40]. Importantly, whilst the AR is the principal transcription factor driving all stages of prostate cancer development, its target genes cannot be easily inferred by coexpression with the AR in contrast to ERG relative to ERG target genes. Target genes for HOXC6, SIM2 and DLX1 are less well defined in prostate cancers but given the presence of ERG and AR target genes within this geneset it is highly likely that they also contribute, being transcription factors, to the expression of some of these genes. A more systematic understanding of the interplay between these transcription factors and dependent gene networks will emerge in future studies. This will require targeting the expression of the transcription factors in experimental model systems and profiling concomitant changes in transcription factor recruitment, chromatin architecture and gene expression.

In the interim, however, it was possible to infer co-dependency based on co-clustering of genes in clinical samples. We did so in two additional datasets, an exon-array dataset generated by Taylor *et al.* and a transcriptomic dataset generated for prostate cancer through high-throughput sequencing by the The Cancer Genome Atlas (TCGA) (Figure 3). In both datasets, we were able to firstly reconfirm the ability of these 33 genes to discriminate between localised prostate cancer and benign tissue samples (Figure 3). Secondly, ERG co-clustered within these 33 genes with bona fide target genes such as B3GAT1 and PLA1A corroborating a contribution at least from ERG to this prostate cancer-specific overexpression signature [39]. Intriguingly, another transcription factor, DLX1, also co-clustered with ERG raising the possibility of a transcription factor hierarchy in which early emergence of an ERG gene fusion may trigger aberrant expression of other developmental transcription factors.

Figure 3 Heatmaps confirming the clustering ability of the 33-gene signature in a prostate cancer exon-array dataset (Taylor *et al.*) and the TCGA RNA-seq dataset for prostate cancer (TCGA-PRAD). The 33-gene signature was applied to two independent datasets, **A.** a prostate cancer exon-array dataset (Taylor *et al.*), and **B.** TCGA RNA-seq dataset for prostate cancer (TCGA-PRAD). Expression values were log2 transformed, normalized for high level mean and variance and hierarchically clustered using Euclidian distance. Genes are rows and samples are columns. The colour coded bars indicate expression values and the clinical grouping for each sample as defined in the keys.

Currently prostate-specific antigen (PSA)/kallikrein 3 (KLK3) is the most widely used protein biomarker for prostate cancer. The androgen receptor (AR) is the most significant transcription factor driving prostate cancer, but is also expressed at high levels in untransformed luminal epithelial cells and therefore is predominantly used as a transcript biomarker associated with metastatic disease and concomitant copy-number amplification [41]. Gene fusions have been detected which significantly elevate the transcript levels of ETS transcription factors and the most prevalent example in prostate cancer is the TMPRSS2-ERG fusion [33]. Detection of the fusion has been reported in biological fluids including urine samples [42].

To assess the performance of the 33-gene signature in comparison to KLK3/PSA, AR and ERG we interrogated an additional independent expression array dataset generated by Grasso *et al.,* and consisting of benign tissue, localised prostate cancer and metastatic cases [43]. This dataset was generated using two different array platforms on distinct sets of samples (Methods section). Cysteine-rich secretory protein-3 (CRISP3) was excluded from the signature due to missing values in the datasets for this gene and prostate cancer antigen 3

(PCA3) was not represented on the arrays leaving a 31-gene signature for evaluation. In the first phase of the signature evaluation we assessed the weighted contribution of each gene in the signature using a logistical regression model on a training dataset consisting of the samples profiled on an Agilent oligo microarray platform. We then used the samples profiled on a second platform, the 4x44K Agilent microarray to evaluate the performance of the signature and compared this to KLK3, ERG and AR. Three pairwise sample comparisons were undertaken - benign versus localised prostate cancer, benign versus metastatic cases and localised prostate cancers versus metastatic cases. Whilst all three transcripts and the signature discriminated between metastatic samples and benign tissue with good specificity and sensitivity as reflected in an area-under-the-curve (AUC) ranging from 0.83 for the AR to 0.99 for the signature, only the signature provided an AUC of $\geq$ 0.95 for all three pairwise comparisons (Table 9). Since both the AR and KLK3 are expressed in both untransformed prostate cells and prostate cancer it is perhaps not surprising that neither yielded an AUC of >0.65 in discriminating between localised prostate cancer and benign tissue samples. By contrast ERG expression is driven by a cancer-associated gene fusion and the AUC was 0.81 (Table 9). AR is amplified and overexpressed in metastatic prostate cancers and this likely explains the higher AUC for this marker (0.84) in discriminating metastatic cases from localised prostate cancers [41]. KLK3/PSA was also higher, 0.87, in this context. ERG by contrast whilst consistently overexpressed in the majority of localised prostate cancers is of variable utility as a prognostic marker according to the study cohort examined associating variously positively or negatively with progression and metastasis [44-46]. In our evaluation the AUC for ERG in discriminating localised prostate cancer from metastatic cases was 0.61, performing more poorly than as a discriminator of localised prostate cancer from benign tissue samples. The AUC differences between the markers and the signature in each pairwise comparison of the sample sets was also visualised in receiver operating characteristic (ROC) curves (Figure 4). These comparisons highlight the importance of using a multi-gene signature since no single gene provides robust discrimination at all stages of the disease, no doubt reflecting changes in the underlying biological drivers during disease progression. We provide in addition AUC values for each individual gene and array probe for each of the pairwise sample comparisons in the test set (Additional file 12: Table S10) and the validation set (Additional file 13: Table S11). Although beyond the scope of this paper we hope that this will assist in further evaluation of the signature by researchers in the field.

Table 9 Comparison of the performance of a 31-gene signature with ERG, AR and KLK3 in discriminating between benign tissue, localised prostate cancer and metastatic disease

	id	agilent_probe	gene	auc.benign.vs.local	auc.benign.vs.met	auc.local.vs.met
1	10308	A_23_P334538	KLK3	0.85625	0.578125	0.55
2	18945	A_24_P108401	KLK3	0.5875	0.8203125	0.8375
3	24972	A_24_P344510	KLK3	0.6875	0.8359375	0.975
4	9045	A_23_P301414	ERG	1	0.75	0.666666666666667
5	15382	A_23_P57323	ERG	0.76875	0.5703125	0.5875
6	31809	A_24_P922378	ERG	0.713333333333333	0.683333333333333	0.5125
7	1039	A_23_P113111	AR	0.51875	0.75	0.75
8	18899	A_24_P106297	AMACR	0.98125	0.7890625	0.8
9	5760	A_23_P1833	B3GAT1	0.80625	0.5703125	0.7375
10	31242	A_24_P914625	BEND4	0.9	0.6875	0.675
11	2222	A_23_P128304	BICD1	0.90625	0.776785714285714	0.542857142857143
12	31375	A_24_P916586	BICD1	0.666666666666667	0.573333333333333	0.6
13	38999	A_32_P472968	BICD1	0.8125	0.5625	0.825
14	31242	A_24_P914625	BEND4	0.9	0.6875	0.675
15	14293	A_23_P45786	COL9A2	0.84375	0.578125	0.725
16	13511	A_23_P419760	CRISP3			
17	12893	A_23_P403466	DLX1	0.71875	0.9453125	0.8
18	34905	A_32_P142818	DLX1	0.955555555555555	1	0.571428571428571
19	18202	A_23_P93058	DNAH5	0.944444444444444	0.94	0.555555555555556
20	19295	A_24_P120462	DNAH5	0.683333333333333	0.625	0.625
21	24667	A_24_P333461	DNAH5	0.555555555555556	0.5	0.611111111111111
22	26067	A_24_P388786	DNAH5	0.95	0.785714285714286	0.657142857142857
23	7096	A_23_P213424	ENC1	0.725	0.9453125	0.8
24	29137	A_24_P69095	ENC1	0.74375	0.517857142857143	0.657142857142857
25	18033	A_23_P91081	EPCAM	0.9625	0.78125	0.575
26	9045	A_23_P301414	ERG	1	0.75	0.666666666666667
27	15382	A_23_P57323	ERG	0.76875	0.5703125	0.5875
28	31809	A_24_P922378	ERG	0.713333333333333	0.683333333333333	0.5125
29	13214	A_23_P412214	RAP1GAP2	0.727272727272727	0.563636363636364	0.666666666666667
30	31091	A_24_P911927	RAP1GAP2	0.6	0.6	0.555555555555556

31	18139	A_23_P9232	GCNT1	0.975	0.705357142857143	0.7
32	5181	A_23_P16523	GDF15	0.7625	0.59375	0.675
33	7901	A_23_P250444	GJB1	0.86875	0.6328125	0.7375
34	26494	A_24_P404840	GJB1	0.85625	0.642857142857143	0.714285714285714
35	11669	A_23_P370666	GLYATL1	0.73125	0.53125	0.75
36	7727	A_23_P2344	HOXC6	0.785714285714286	0.958333333333333	0.857142857142857
37	147	A_23_P101806	HPN	0.925	0.75	0.5125
38	13011	A_23_P406782	HPN	0.90625	0.703125	0.8
39	6231	A_23_P203933	ITPR2	0.84	0.761904761904762	0.942857142857143
40	9378	A_23_P310	MARCKSL1	0.93125	0.921875	0.6625
41	2744	A_23_P13442	MICAL2	0.716049382716049	0.62962962962963	0.518518518518518
42	7851	A_23_P24843	MICAL2	0.8	0.5625	0.75
43	7314	A_23_P215956	MYC	0.8125	0.7578125	0.9125
44	20716	A_24_P178011	MYC	0.626666666666667	0.733333333333333	0.857142857142857
45	25948	A_24_P38363	MYC	0.5	1	1
46	3126	A_23_P139327	OR51E2	0.85	0.645833333333333	0.683333333333333
47	22188	A_24_P235756	OR51E2	0.88125	0.5625	0.8375
48	8971	A_23_P29816	PLA1A	0.925	0.640625	0.725
49	23649	A_24_P294408	PLA1A	0.9625	0.580357142857143	0.885714285714286
50	15416	A_23_P5778	RAB17	0.928571428571429	0.830357142857143	0.675
51	7973	A_23_P251387	REPS2	0.9875	0.78125	0.8125
52	33496	A_32_P100109	REPS2	0.94375	0.7890625	0.9625
53	6876	A_23_P211110	SIM2	0.914285714285714	0.803571428571429	0.7375
54	9061	A_23_P301886	SIM2	0.8	0.916666666666667	0.5
55	15035	A_23_P52939	SLC43A1	0.925	0.671875	0.6625
56	22855	A_24_P260443	THBS4	0.8125	0.8046875	0.9375
57	1849	A_23_P123503	TRIB1	0.88125	0.642857142857143	0.528571428571429
58	22636	A_24_P252497	TRIB1	0.775	0.578125	0.825
59	4805	A_23_P160460	UAP1	0.9625	0.7265625	0.9625

Data were downloaded from Grasso *et al.*,. ROC statistics were computed in an evaluation sample set having established the weighting for genes in the signature using logistic regression in a test sample set. We report the area under the curve (AUC) for each transcript and for the signature for each of three pairwise comparisons as generated using the R package ROCR.

Figure 4 Receiver operating characteristic (ROC) curves for discrimination between localised prostate cancer and benign cases, metastatic and benign cases and metastatic and prostate cancers using a 31-gene signature (row 1), AR (row 2), ERG (row 3) and KLK3 (row 4).

Conclusions

In conclusion, in this study we have used a multi-step approach to refine gene signatures derived from diverse transcript detection platforms and sample types in order to arrive at a robust gene signature able to discriminate between PCa and benign tissue (Figure 5). This is the first time that this has been attempted and demonstrates that value exists in transcript signatures generated from amongst the earliest microarray studies right through to high-throughput sequencing. In brief, beginning with a small expression array dataset consisting of 13 macrodissected samples, we have been able to derive gene signatures capable of subclustering localised PCa and metastases in a larger microdissected and a multi-cancer dataset (Figure 5). This highlights that there are valuable gene transcript signatures that can be robust despite cellular heterogeneity in PCa and the evolution of transcript detection technologies. In addition, we have discovered that gene transcripts that are significantly overexpressed within these signatures are also overexpressed in much more recently acquired exon-array and sequence-based TCGA data, transcript detection platforms that were unavailable when the Varambally, Tomlins and Ramaswamy studies were undertaken (Figure 5). Finally, we have evaluated the performance of these transcripts as a signature in discriminating between benign tissue samples, localised PCa and metastatic disease in an additional dataset generated by Grasso *et al.* ROC curves reveal that the signature exceeds the performance of ERG, KLK3 or the AR as a classifier. Intriguingly, one third of these genes are glycosylating enzymes and transcription factors. PCa is significantly driven by a transcription factor, the AR, but there is increasing evidence of contributions by others and of interplays between them and indeed our signature does not include the AR itself. However, it includes both established examples (MYC and ERG) but also others that have so far been less studied (SIM2, DLX1 and HOXC6). Mechanistically, future work will investigate this transcriptional co-dependency in more detail and clinically these signatures will be further evaluated in clinical cohorts.

Figure 5 Workflow for the identification of robust gene signatures and gene sets for clustering prostate cancer cases. In step 1, we identified all statistically significant differentially expressed Affymetrix array probes in a small dataset consisting of 13 macrodissected clinical samples encompassing localised benign prostatic hyperplasia, localised prostate cancer and metastatic disease (GSE3325). We then generated gene signatures from these based on gene coexpression at varying stringency thresholds. These gene signatures were then applied to two additional datasets, a microdissected dataset (Tomlins et *al.*) and a multi-tissue site cancer and metastatic dataset (Ramaswamy *et al.*). A large number of the coexpression gene signatures clustered localised prostate cancers from metastatic disease and prostate metastases from metastases at other organ sites. The most compact gene signature able to do so consisted of 71 genes and we assessed its expression pattern in two additional datasets, an exon-array dataset (Taylor *et al.*) and in a RNA-sequenced dataset (TCGA-PRAD). Few of the genes in the significant coexpression gene signatures were overexpressed genes in localised prostate cancers. In the second phase of the study, we abstracted all of the overexpressed genes and refined this down to a set of 33 genes based on significant overexpression in additional publicly available prostate cancer microarray datasets housed within the Oncomine database. These genes also effectively clustered benign versus cancer cases in an exon-array dataset (Taylor *et al.*) an expression microarray dataset (Grasso *et al.*) and a RNA-sequenced dataset (TCGA-PRAD). In conclusion, it is possible to generate gene classifiers of clinical prostate cancer from a small dataset of macrodissected samples with the capacity to classify larger sequenced and microdissected datasets based on clinical characteristics.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

I.G.M. conceived the study and wrote the paper. P.E. implemented the Ward agglomerative method and analysed the Varambally, Tomlins and Ramaswamy datasets and also wrote the paper. S.J.B. analysed the Taylor and TCGA datasets and wrote the paper. V.Z. evaluated the performance of the gene signature in the Grasso dataset. All authors read and approved the final manuscript.

Acknowledgements

I.G.M. is supported by funding from the Norwegian Research Council, Helse Sor-Ost and the University of Oslo through the Centre for Molecular Medicine (Norway), which is the part of the Nordic EMBL (European Molecular Biology Laboratory) partnership and also supported by Oslo University Hospitals. I.G.M. is also supported by the Norwegian Cancer Society and by EU FP7 funding. I.G.M. holds a visiting scientist position with Cancer Research UK through the Cambridge Research Institute and a Senior Visiting Research Fellowship with Cambridge University through the Department of Oncology. S.J.B is funded by the Norwegian Cancer Society and Molecular Life Sciences at the University of Oslo. V. Z. is support by the Centre for Molecular Medicine (Norway) and the Psychosis Research Centre at the Institute for Clinical Medicine at Oslo university Hospitals. P. E. is supported by Cancer Research UK.

References

1. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci U S A* 2004, **101:**9309?9314.

2. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36:**1090?1098.

3. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, *et al*: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.** *Proc Natl Acad Sci U S A* 2006, **103:**17402?17407.

4. Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, Masys D, Arden K, Goodison S, McClelland M, *et al*: **In silico dissection of cell-type-associated patterns of gene expression in prostate cancer.** *Proc Natl Acad Sci U S A* 2004, **101:**615?620.

5. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, *et al*: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39:**41?51.

6. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat Genet* 2003, **33:**49?54.

7. Hoerl AE, Kennard RW: **Ridge regression: biased estimation for nonorthogonal problems.** *Technometrics* 1970, **12:**55?67.

8. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, *et al*: **Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression.** *Cancer Cell* 2005, **8:**393?406.

9. Ward JH Jr: **Hierarchical grouping to optimize an objective function.** *J Am Stat Assoc* 1963, **58:**236?244.

10. McQuitty LL: **Similarity analysis by reciprocal pairs for discrete and continuous data.** *Educ Psychol Measure* 1966**:**825-831.

11. Shannon CE: **A mathematical theory of communication.** *Bell Syst Tech J* 1948, **27:**379?423.

12. Chen B-S, Li C-W: **On the interplay between entropy and robustness of gene regulatory networks.** *Entropy* 2010, **12:**1071?1101.

13. Furlanello C, Serafini M, Merler S, Jurman G: **Entropy-based gene ranking without selection bias for the predictive classification of microarray data.** *BMC Bioinformatics* 2003, **4:**54.

14. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, Mesher D, Speights VO, Stankiewicz E, Foster CS, *et al*: **Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study.** *Lancet Oncol* 2011, **12:**245?255.

15. Fukuma Y, Matsui H, Koike H, Sekine Y, Shechter I, Ohtake N, Nakata S, Ito K, Suzuki K: **Role of squalene synthase in prostate cancer risk and the biological aggressiveness of human prostate cancer.** *Prostate Cancer Prostatic Dis* 2012, **15:**339?345.

16. Cuzick J, Berney DM, Fisher G, Mesher D, Moller H, Reid JE, Perry M, Park J, Younus A, Gutin A, *et al*: **Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort.** *Br J Cancer* 2012, **106:**1095?1099.

17. Yuan X, Cai C, Chen S, Yu Z, Balk SP: **Androgen receptor functions in castration-resistant prostate cancer and mechanisms of resistance to new agents targeting the androgen axis.** *Oncogene* 2014, **33:**2815?2825.

18. Jenkins RB, Qian J, Lieber MM, Bostwick DG: **Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization.** *Cancer Res* 1997, **57:**524?531.

19. Ellwood-Yen K, Graeber TG, Wongvipat J, Iruela-Arispe ML, Zhang J, Matusik R, Thomas GV, Sawyers CL: **Myc-driven murine prostate cancer shares molecular features with human prostate tumors.** *Cancer Cell* 2003, **4:**223?238.

20. Mikaelsson E, Danesh-Manesh AH, Luppert A, Jeddi-Tehrani M, Rezvany MR, Sharifian RA, Safaie R, Roohi A, Osterborg A, Shokri F, *et al*: **Fibromodulin, an extracellular matrix protein: characterization of its unique gene and protein expression in B-cell chronic lymphocytic leukemia and mantle cell lymphoma.** *Blood* 2005, **105:**4828?4835.

21. Vallat L, Magdelenat H, Merle-Beral H, Masdehors P, Potocki De Montalk G, Davi F, Kruhoffer M, Sabatier L, Orntoft TF, Delic J: **The resistance of B-CLL cells to DNA damage-induced apoptosis defined by DNA microarrays.** *Blood* 2003, **101:**4598?4606.

22. Levens E, Luo X, Ding L, Williams RS, Chegini N: **Fibromodulin is expressed in leiomyoma and myometrium and regulated by gonadotropin-releasing hormone analogue therapy and TGF-beta through Smad and MAPK-mediated signalling.** *Mol Hum Reprod* 2005, **11:**489?494.

23. Di Vizio D, Morello M, Sotgia F, Pestell RG, Freeman MR, Lisanti MP: **An absence of stromal caveolin-1 is associated with advanced prostate cancer, metastatic disease and epithelial Akt activation.** *Cell Cycle* 2009, **8:**2420?2424.

24. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, *et al*: **Integrative genomic profiling of human prostate cancer.** *Cancer Cell* 2010, **18:**11?22.

25. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform.** *Neoplasia* 2004, **6:**1?6.

26. Itkonen HM, Engedal N, Babaie E, Luhr M, Guldvik IJ, Minner S, Hohloch J, Tsourlakis MC, Schlomm T, Mills IG: **UAP1 is overexpressed in prostate cancer and is protective against inhibitors of N-linked glycosylation.** *Oncogene* 2014.

27. Hagisawa S, Ohyama C, Takahashi T, Endoh M, Moriya T, Nakayama J, Arai Y, Fukuda M: **Expression of core 2 beta1,6-N-acetylglucosaminyltransferase facilitates prostate cancer progression.** *Glycobiology* 2005, **15:**1016?1024.

28. Ma W, Diep K, Fritsche HA, Shore N, Albitar M: **Diagnostic and prognostic scoring system for prostate cancer using urine and plasma biomarkers.** *Genet Test Mol Biomarkers* 2014, **18:**156?163.

29. Kamigaito T, Okaneya T, Kawakubo M, Shimojo H, Nishizawa O, Nakayama J: **Overexpression of O-GlcNAc by prostate cancer cells is significantly associated with poor prognosis of patients.** *Prostate Cancer Prostatic Dis* 2014, **17:**18?22.

30. Itkonen HM, Minner S, Guldvik IJ, Sandmann MJ, Tsourlakis MC, Berge V, Svindland A, Schlomm T, Mills IG: **O-GlcNAc transferase integrates metabolic pathways to regulate the stability of c-MYC in human prostate cancer cells.** *Cancer Res* 2013, **73:**5277?5287.

31. Gurel B, Iwata T, Koh CM, Jenkins RB, Lan F, Van Dang C, Hicks JL, Morgan J, Cornish TC, Sutcliffe S, *et al*: **Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis.** *Mod Pathol* 2008, **21:**1156?1167.

32. Chou TY, Hart GW, Dang CV: **c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas.** *J Biol Chem* 1995, **270:**18961?18965.

33. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, *et al*: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.** *Science* 2005, **310:**644?648.

34. Deyoung MP, Scheurle D, Damania H, Zylberberg C, Narayanan R: **Down s syndrome-associated single minded gene as a novel tumor marker.** *Anticancer Res* 2002, **22:**3149?3157.

35. Halvorsen OJ, Rostad K, Oyan AM, Puntervoll H, Bo TH, Stordrange L, Olsen S, Haukaas SA, Hood L, Jonassen I, *et al*: **Increased expression of SIM2-s protein is a novel marker of aggressive prostate cancer.** *Clin Cancer Res* 2007, **13:**892?897.

36. Koh CM, Gurel B, Sutcliffe S, Aryee MJ, Schultz D, Iwata T, Uemura M, Zeller KI, Anele U, Zheng Q, *et al*: **Alterations in nucleolar structure and gene expression programs in prostatic neoplasia are driven by the MYC oncogene.** *Am J Pathol* 2011, **178:**1824?1834.

37. Wang L, Li Y, Yang X, Yuan H, Li X, Qi M, Chang YW, Wang C, Fu W, Yang M, *et al*: **ERG-SOX4 interaction promotes epithelial-mesenchymal transition in prostate cancer cells.** *Prostate* 2014, **74:**647?658.

38. Leshem O, Madar S, Kogan-Sakin I, Kamer I, Goldstein I, Brosh R, Cohen Y, Jacob-Hirsch J, Ehrlich M, Ben-Sasson S, *et al*: **TMPRSS2/ERG promotes epithelial to mesenchymal transition through the ZEB1/ZEB2 axis in a prostate cancer model.** *PLoS One* 2011, **6:**e21650.

39. Ribeiro FR, Paulo P, Costa VL, Barros-Silva JD, Ramalho-Carvalho J, Jeronimo C, Henrique R, Lind GE, Skotheim RI, Lothe RA, Teixeira MR: **Cysteine-rich secretory protein-3 (CRISP3) is strongly up-regulated in prostate carcinomas with the TMPRSS2-ERG fusion gene.** *PLoS One* 2011, **6:**e22317.

40. Itkonen HM, Mills IG: **N-linked glycosylation supports cross-talk between receptor tyrosine kinases and androgen receptor.** *PLoS One* 2013, **8:**e65016.

41. Visakorpi T, Hyytinen E, Koivisto P, Tanner M, Keinanen R, Palmberg C, Palotie A, Tammela T, Isola J, Kallioniemi OP: **In vivo amplification of the androgen receptor gene and progression of human prostate cancer.** *Nat Genet* 1995, **9:**401?406.

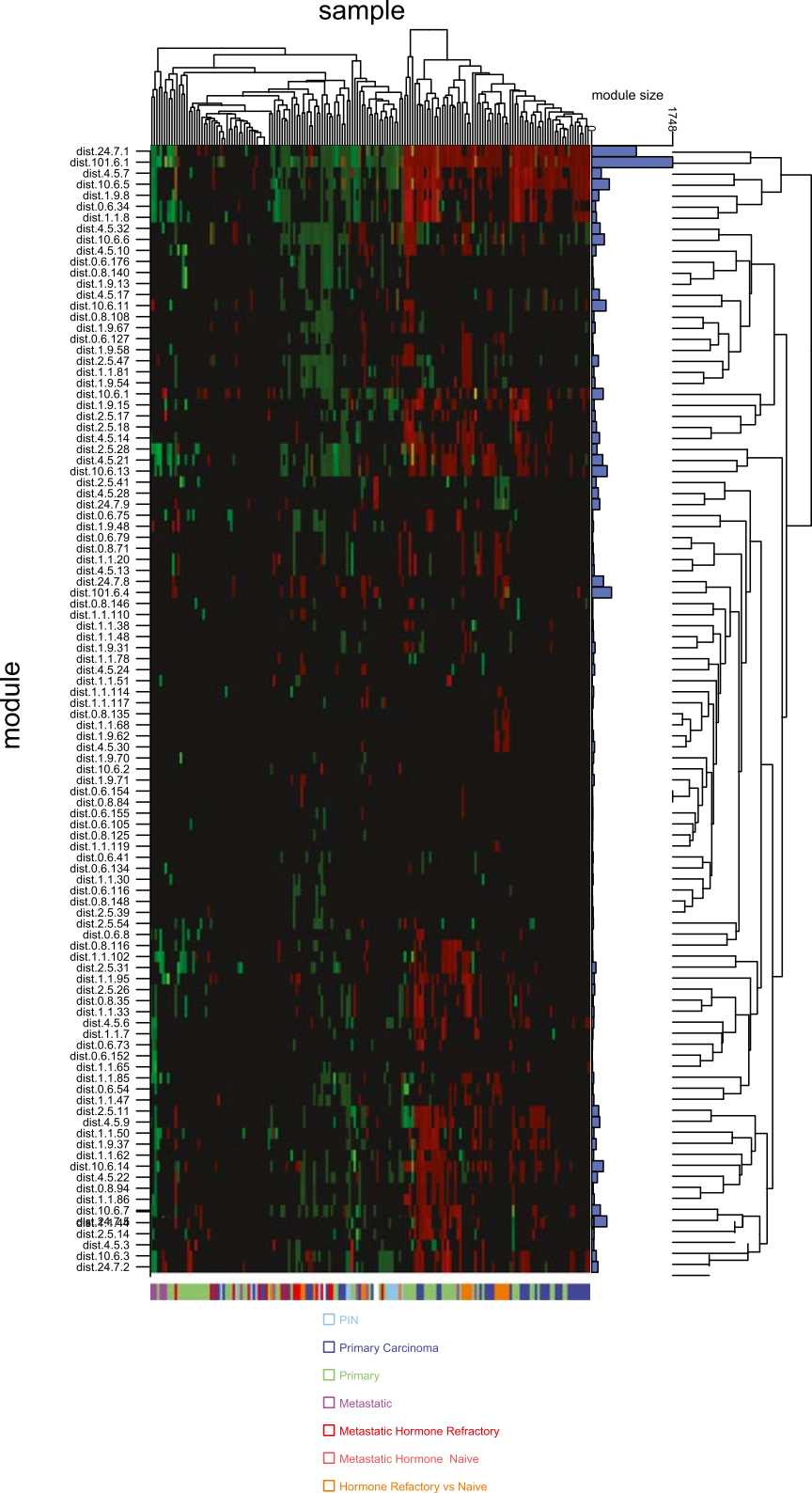
42. Tomlins SA, Aubin SM, Siddiqui J, Lonigro RJ, Sefton-Miller L, Miick S, Williamsen S, Hodge P, Meinke J, Blase A, *et al*: **Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA.** *Sci Transl Med* 2011, **3:**94ra72.

43. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC, *et al*: **The mutational landscape of lethal castration-resistant prostate cancer.** *Nature* 2012, **487:**239?243.

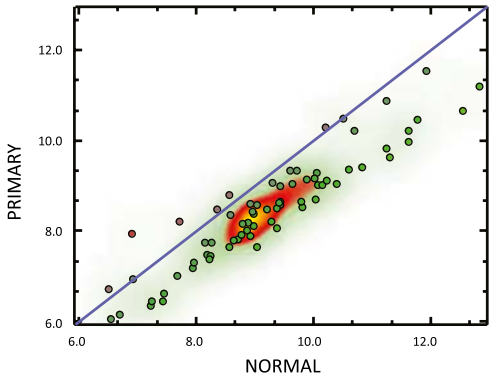
44. Fleischmann A, Saramaki OR, Zlobec I, Rotzer D, Genitsch V, Seiler R, Visakorpi T, Thalmann GN: **Prevalence and prognostic significance of TMPRSS2-ERG gene fusion in lymph node positive prostate cancers.** *Prostate* 2014, **74:**1647?1654.

45. Nam RK, Sugar L, Wang Z, Yang W, Kitching R, Klotz LH, Venkateswaran V, Narod SA, Seth A: **Expression of TMPRSS2:ERG gene fusion in prostate cancer cells is an important prognostic factor for cancer progression.** *Cancer Biol Ther* 2007, **6:**40?45.

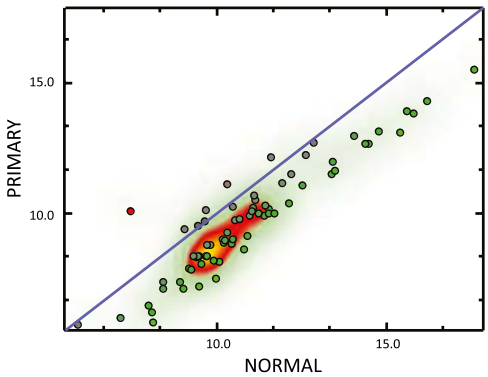
46. Saramaki OR, Harjula AE, Martikainen PM, Vessella RL, Tammela TL, Visakorpi T: **TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis.** *Clin Cancer Res* 2008, **14:**3395?3400.


sample
module size
1748
module
dist.24.7.1
dist.101.6.1
dist.4.5.7
dist.10.6.5
dist.1.9.8
dist.0.6.34
dist.1.1.8
dist.4.5.32
dist.10.6.6
dist.4.5.10
dist.0.6.176
dist.0.8.140
dist.1.9.13
dist.4.5.17
dist.10.6.11
dist.0.8.108
dist.1.9.67
dist.0.6.127
dist.1.9.58
dist.2.5.47
dist.1.1.81
dist.1.9.54
dist.10.6.1
dist.1.9.15
dist.2.5.17
dist.2.5.18
dist.4.5.14
dist.2.5.28
dist.4.5.21
dist.10.6.13
dist.2.5.41
dist.4.5.28
dist.24.7.9
dist.0.6.75
dist.1.9.48
dist.0.6.79
dist.0.8.71
dist.1.1.20
dist.4.5.13
dist.24.7.8
dist.101.6.4
dist.0.8.146
dist.1.1.110
dist.1.1.38
dist.1.1.48
dist.1.9.31
dist.1.1.78
dist.4.5.24
dist.1.1.51
dist.1.1.114
dist.1.1.117
dist.0.8.135
dist.1.1.68
dist.1.9.62
dist.4.5.30
dist.1.9.70
dist.10.6.2
dist.1.9.71
dist.0.6.154
dist.0.8.84
dist.0.6.155
dist.0.6.105
dist.0.8.125
dist.1.1.119
dist.0.6.41
dist.0.6.134
dist.1.1.30
dist.0.6.116
dist.0.8.148
dist.2.5.39
dist.2.5.54
dist.0.6.8
dist.0.8.16
dist.1.1.102
dist.2.5.31
dist.1.1.95
dist.2.5.26
dist.0.8.35
dist.1.1.33
dist.4.5.6
dist.1.1.7
dist.0.6.73
dist.0.6.152
dist.1.1.65
dist.1.1.85
dist.0.6.54
dist.1.1.47
dist.2.5.11
dist.4.5.9
dist.1.1.50
dist.1.9.37
dist.1.1.62
dist.10.6.14
dist.4.5.22
dist.0.8.94
dist.1.1.86
dist.10.6.7
dist.24.7.5
dist.2.5.14
dist.4.5.3
dist.10.6.3
dist.24.7.2
PIN
Primary Carcinoma
Primary
Metastatic
Metastatic Hormone Refractory
Metastatic Hormone Naive
Hormone Refactory vs Naive

A Taylor *et al.*



B TCGA-PRAD

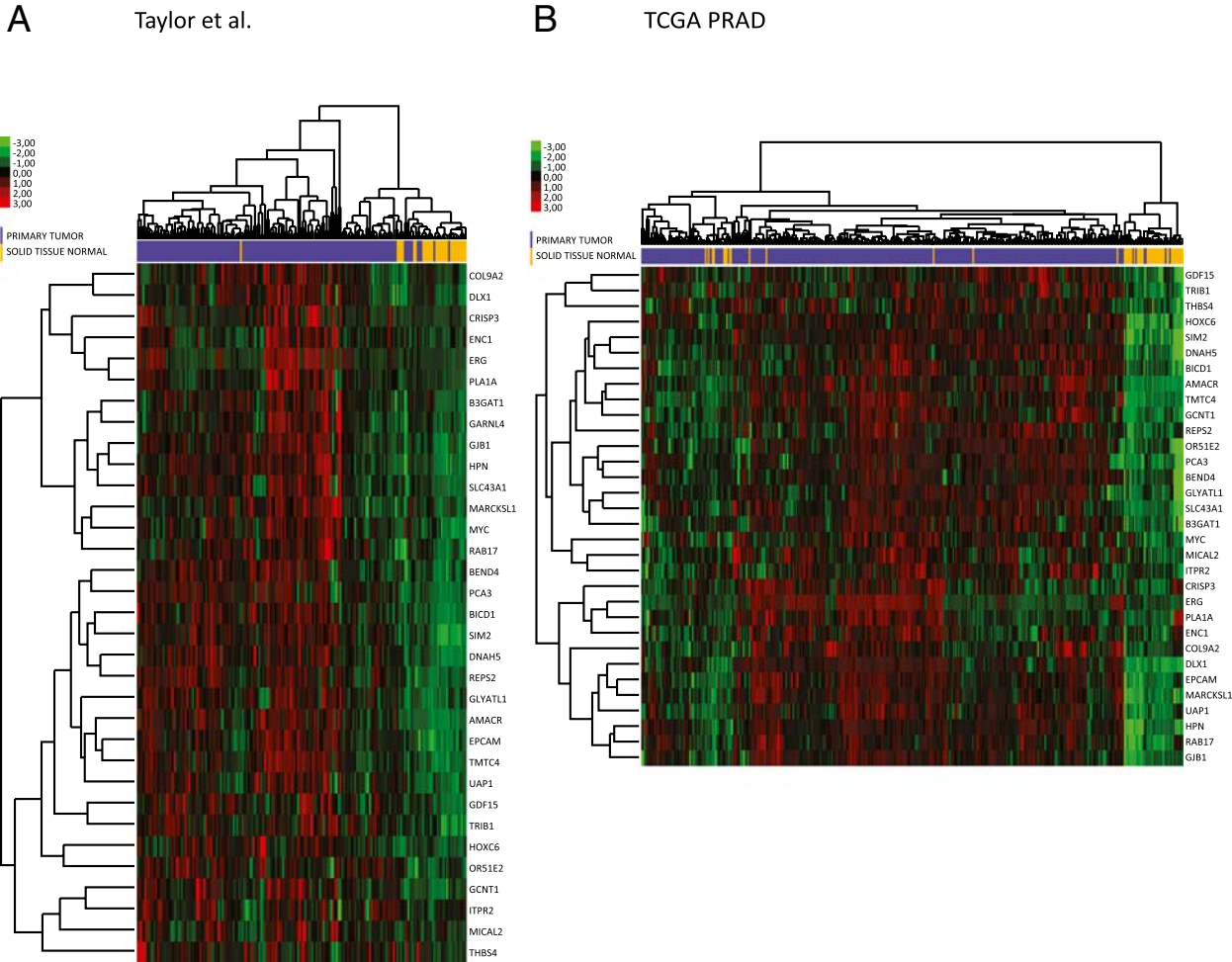


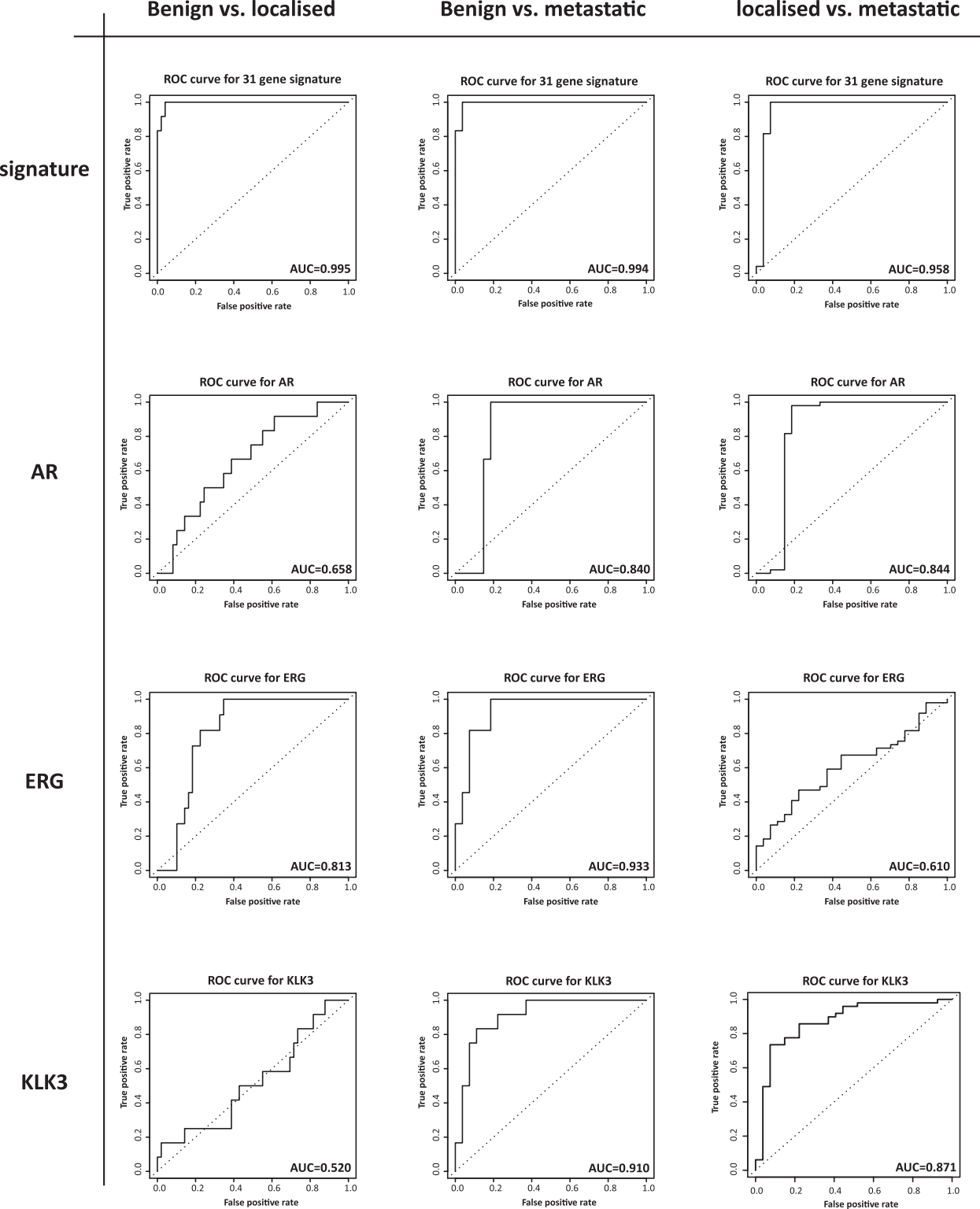
SYMBOL	NORMAL	PRIMARY	FOLD CHANGE
GLYATL1	6.906774	7.917742	2.015263
MYC	7.707574	8.185126	1.392379
FMOD	8.560921	8.76869	1.154901
CCL4	6.514742	6.676367	1.118547
ZNF615	8.361406	8.461288	1.071686
MCL1	10.20404	10.26873	1.045861
CYP39A1	6.920818	6.894722	-1.01825
FOS	10.50548	10.46167	-1.03083
EBPL	8.575874	8.323491	-1.19117
FAM36A	9.301839	9.045351	-1.19457
CYR61	9.596716	9.309407	-1.22036
ADARB1	8.918336	8.577855	-1.26618
C3orf64	11.23722	10.86437	-1.29491
IGF1	11.92109	11.51532	-1.3248
DPP4	9.722254	9.315641	-1.32557
FOXF1	8.153576	7.719392	-1.35115
SIN3A	9.432331	8.954617	-1.39254
IRS1	9.029387	8.544991	-1.399
MXRA7	10.68732	10.19753	-1.40424
PAGE4	6.544602	6.022728	-1.43582
GBP1	8.261975	7.721639	-1.45431
SFRP1	8.965537	8.395006	-1.48507
PLN	6.697379	6.117262	-1.49497
LOC728264	8.977178	8.356482	-1.53762
SLC22A3	9.639426	9.015032	-1.54156
OSR2	8.790582	8.122536	-1.58892
CLIC6	7.684455	6.989104	-1.61928
PABPC4L	7.962172	7.264653	-1.62171
MRGPRF	8.87798	8.158198	-1.64693
CLIP4	8.191268	7.455136	-1.6657
PDE4D	9.214379	8.457332	-1.69003
FHL2	9.8855	9.111159	-1.71041
LTF	10.05994	9.272211	-1.72635
ATP1A2	9.423973	8.63473	-1.72817
OGN	7.950126	7.146403	-1.7456

SYMBOL	NORMAL	PRIMARY	FOLD CHANGE
SSPN	9.409508	8.603026	-1.74894
GLI3	8.241894	7.417086	-1.7713
RGN	7.244353	6.415509	-1.77626
SEMA3C	10.01368	9.146269	-1.82439
TSPAN2	8.227478	7.358064	-1.82692
COPZ2	8.860626	7.988839	-1.82993
CAMK2G	9.425827	8.553151	-1.83106
EYA4	7.453188	6.575563	-1.83735
PRICKLE2	8.647058	7.767708	-1.83955
TCF21	8.774418	7.893442	-1.84162
PI15	9.37299	8.484863	-1.85077
KANK1	8.978575	8.081426	-1.86238
PDZRN4	7.224421	6.325374	-1.86483
RND3	8.559543	7.629499	-1.90533
HLF	8.724327	7.784206	-1.91869
ANGPT1	7.425034	6.409305	-2.02192
ALDH1A2	8.928527	7.852115	-2.10879
CDC42EP3	10.07616	8.988163	-2.12579
CFL2	9.28068	8.18929	-2.13079
MSRB3	10.23365	9.099704	-2.19457
WFDC1	9.791128	8.625923	-2.24265
KCNMB1	10.15205	8.983879	-2.24727
MYLK	10.59709	9.335036	-2.39837
MAOB	9.802727	8.508808	-2.45193
LOC572558	9.377546	8.034269	-2.53727
MYL9	11.77468	10.428	-2.54326
MYOF	10.04425	8.661866	-2.607
LMOD1	10.39851	9.003799	-2.62936
MYOCD	9.031196	7.622591	-2.6548
TNS1	11.23702	9.811136	-2.68679
DES	11.62794	10.19113	-2.70722
TPM2	10.82443	9.37415	-2.73261
ACTG2	12.82488	11.17131	-3.14611
NEFH	11.62854	9.949989	-3.20107
CNN1	11.30499	9.603444	-3.2525
MYH11	12.53658	10.64449	-3.71171

SYMBOL	NORMAL	PRIMARY	FOLD CHANGE
GLYATL1	7.457804	10.02644	5.932473
MYC	10.29389	11.08235	1.727225
FMOD	11.61548	12.10059	1.399688
ZNF615	9.688937	10.08872	1.319305
CYP39A1	9.060082	9.352686	1.224849
EBPL	9.44122	9.488349	1.033206
ADARB1	9.629741	9.639631	1.006879
MCL1	12.87639	12.66382	-1.15875
CCL4	5.916898	5.682089	-1.17675
FAM36A	10.47209	10.19277	-1.21362
DPP4	12.62986	12.19765	-1.3493
SIN3A	11.09953	10.61601	-1.39815
SFRP1	11.13197	10.46435	-1.58845
CYR61	12.18708	11.44959	-1.66727
SEMA3C	11.92836	11.08574	-1.79331
IRS1	10.55508	9.696492	-1.81326
PI15	11.07392	10.14147	-1.90852
PDE4D	10.67238	9.709418	-1.94931
OSR2	9.700553	8.731292	-1.95784
TCF21	9.305067	8.324244	-1.97359
SLC22A3	11.05038	10.04227	-2.01127
C3orf64	9.821982	8.756526	-2.09283
COPZ2	8.422251	7.329426	-2.13291
IGF1	10.29856	9.195401	-2.14825
CAMK2G	10.97351	9.86686	-2.15345
FOS	14.04549	12.93219	-2.16339
LOC728264	9.450622	8.322782	-2.18531
GBP1	9.464016	8.307469	-2.22923
FOXF1	9.422833	8.264921	-2.23134
PABPC4L	7.161422	5.957948	-2.30294
MRGPRF	10.16807	8.964191	-2.30357
KANK1	11.4424	10.23472	-2.30966
SSPN	10.24439	8.931919	-2.48367
FHL2	11.24463	9.931912	-2.48409
GLI3	9.16998	7.849303	-2.49783
PAGE4	8.416857	7.041602	-2.59414

SYMBOL	NORMAL	PRIMARY	FOLD CHANGE
CLIP4	9.67412	8.298259	-2.59523
CDC42EP3	11.52193	10.13509	-2.61505
OGN	10.2151	8.825615	-2.61985
HLF	9.724102	8.311322	-2.6625
TSPAN2	9.258912	7.781335	-2.78481
NEFH	13.43626	11.93944	-2.8222
WFDC1	10.46986	8.946525	-2.87454
MXRA7	12.54074	11.00498	-2.89942
CFL2	11.39027	9.840245	-2.92821
PLN	9.562167	8.00411	-2.94457
MAOB	11.52369	9.95599	-2.96432
CLIC6	8.907655	7.336251	-2.97194
RGN	8.004386	6.411789	-3.01592
RND3	10.43499	8.778803	-3.15182
TPM2	14.7658	13.0901	-3.19475
MYL9	15.60606	13.84934	-3.37929
PRICKLE2	9.897849	8.139418	-3.3833
MSRB3	11.69072	9.928429	-3.39237
TNS1	14.38974	12.61048	-3.4325
MYOF	12.12388	10.34144	-3.44006
KCNMB1	10.9181	9.0827	-3.56869
CNN1	14.48048	12.61957	-3.63236
LTF	13.49969	11.59508	-3.74407
EYA4	8.077518	6.155506	-3.78951
DES	16.19062	14.25205	-3.83326
LMOD1	13.3857	11.44545	-3.8377
PDZRN4	9.014486	7.061059	-3.87293
ALDH1A2	10.06106	8.094925	-3.90721
ACTG2	15.82486	13.78335	-4.11676
MYH11	17.61345	15.45035	-4.47875
ATP1A2	10.79166	8.578395	-4.63724
MYOCD	9.47278	7.125689	-5.08797
MYLK	15.42531	13.05382	-5.17478
LOC572558	8.136008	5.759487	-5.19283
ANGPT1	9.969429	7.461877	-5.68654


A
Taylor et al.
-3,00
-2,00
-1,00
0,00
1,00
2,00
3,00
PRIMARY TUMOR
SOLID TISSUE NORMAL
COL9A2
DLX1
CRISP3
ENC1
ERG
PLA1A
B3GAT1
GARNL4
GJB1
HPN
SLC43A1
MARCKSL1
MYC
RAB17
BEND4
PCA3
BICD1
SIM2
DNAH5
REPS2
GLYATL1
AMACR
EPCAM
TMTC4
UAP1
GDF15
TRIB1
HOXC6
OR51E2
GCNT1
ITPR2
MICAL2
THBS4
B
TCGA PRAD
-3,00
-2,00
-1,00
0,00
1,00
2,00
3,00
PRIMARY TUMOR
SOLID TISSUE NORMAL
GDF15
TRIB1
THBS4
HOXC6
SIM2
DNAH5
BICD1
AMACR
TMTC4
GCNT1
REPS2
OR51E2
PCA3
BEND4
GLYATL1
SLC43A1
B3GAT1
MYC
MICAL2
ITPR2
CRISP3
ERG
PLA1A
ENC1
COL9A2
DLX1
EPCAM
MARCKSL1
UAP1
HPN
RAB17
GJB1


Benign vs. localised
Benign vs. metastatic
localised vs. metastatic
signature
AR
ERG
KLK3
ROC curve for 31 gene signature
AUC=0.995
AUC=0.994
AUC=0.958
ROC curve for AR
AUC=0.658
AUC=0.840
AUC=0.844
ROC curve for ERG
AUC=0.813
AUC=0.933
AUC=0.610
ROC curve for KLK3
AUC=0.520
AUC=0.910
AUC=0.871
True positive rate
False positive rate

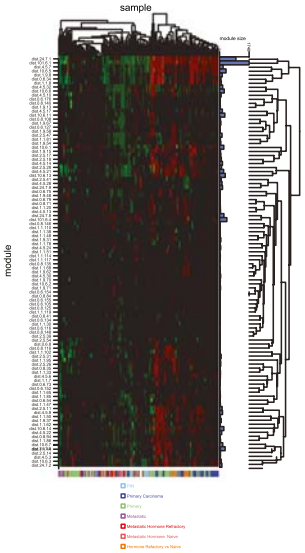
Varambally *et al.* (2005)
whole tissue dataset

112 genes BPH -> primary
2900 genes primary -> metastatic

4 coexpressed gene signatures

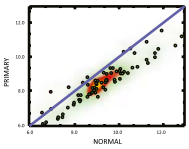
signature 1 (101.6.1)
signature 2 (101.6.2)
signature 3 (101.6.3)
signature 4 (101.6.4)

Tomlins *et al.* (2007)
Laser microdissected

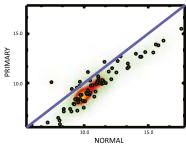


Ramaswamy *et al.* (2003)
multiple tissue sites

Taylor *et al.* (2010)
microarrays



TCGA-PRAD
RNA-seq



A. validation of smallest gene signature (71 genes) in two other datasets

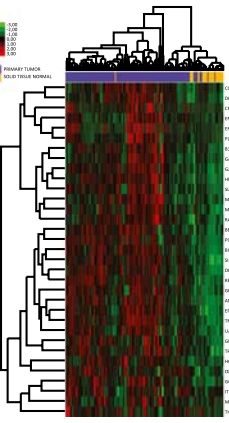
gene signatures capable of clustering primary and metastatic samples.

B. continue with the 97 overexpressed genes in the signatures

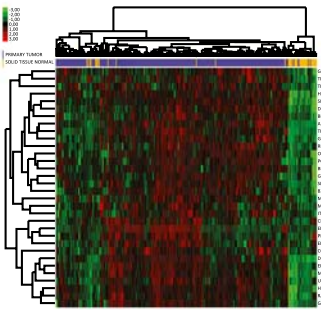
33 genes are in the TOP1% overexpressed in at least three other datasets

A. visualization in two datasets

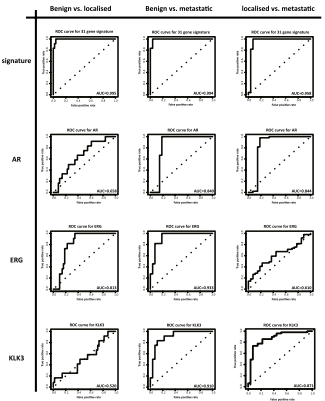
Taylor *et al.*



TCGA PRAD



B. ROC curves in *Grasso et al.*



Additional files provided with this submission:

Additional file 1: Figure S1. W(C) values generated from branch length thresholds plotted against the number of clusters produced for single, average, complete, ward and mcquitty agglomerative hierarchical and divisive clustering methods (245k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s1.zip

Additional file 2: Table S1. Annotated Affymterix probes for differentially expressed genes in localised prostate cancer and metastatic tumours. Genes differentially expressed within dataset GSE3325. Columns 1 and 2 represent Affymetrix probe identifiers and gene symbols, respectively. Columns 3-5 represent mean probe signals for benign, prostate cancer and metastatic samples. Columns 6-9 represent a linear fold change and corresponding p-value for each contrast. Genes are ranked based on overexpression in localised prostate cancer versus benign tissue. Differential expression and associated statistical significance (p-values) have been calculated as described in the methods/supplementary methods section (1403k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s2.xls

Additional file 3: Figure S2. H(X) GO bit values generated from branch length thresholds plotted against the number of clusters produced for single, average, complete, Ward and Mcquitty agglomerative hierarchical and divisive clustering methods (167k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s3.zip

Additional file 4: Table S2. Differentially expressed genes within Varambally dataset (accession number GSE3325) grouped into coexpression signatures using the Ward agglomerative method. Genes are ranked based on overexpression in localised prostate cancer versus benign tissue. Annotation includes the cytogenetic locus for each gene and the gene name. Coexpression modules are defined at multiple branch thresholds. Nine branch thresholds were used. Dist.0.6 represents the most compact modules at the highest stringency threshold and dist.101.6 representing the least compact modules at the lowest stringency threshold. Numbers attributed to each gene within each branch threshold define membership of a distinct coexpression module. With reference to the main text ‘signature 1’ corresponds to 101.6.1, ‘signature 2’ corresponds to 101.6.2, ‘signature 3’ corresponds to 101.6.3 and ‘signature 4’ corresponds to 101.6.4 (1078k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s4.xlsx

Additional file 5: Table S3. Differentially expressed genes comprising signature 1. Genes within signature 1 and sub-signatures defined at more stringent branch thresholds are ranked based on overexpression in localised prostate cancer within dataset GSE3325. Annotation includes the cytogenetic loci, gene names and unigene modules (1085k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s5.xlsx

Additional file 6. Table S4 Differentially expressed genes comprising signature 2. Genes within signature 2 and sub-signatures defined at more stringent branch thresholds are ranked based on overexpression in localised prostate cancer within dataset GSE3325. Annotation includes the cytogenetic loci, gene names and unigene modules (1085k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s6.xlsx

Additional file 7: Table S5. Differentially expressed genes comprising signature 3. Genes within signature 3 and sub-signatures defined at more stringent branch thresholds are ranked based on overexpression in localised prostate cancer within dataset GSE3325. Annotation includes the cytogenetic loci, gene names and unigene modules (1084k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s7.xlsx

Additional file 8: Table S6. Differentially expressed genes comprising signature 4. Genes within signature 4 and sub-signatures defined at more stringent branch thresholds are ranked based on overexpression in localised prostate cancer within dataset GSE3325. Annotation includes the cytogenetic loci, gene names and unigene modules (1082k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s8.xlsx

Additional file 9: Table S7. Fold-change in coexpressed gene signatures derived from Varambally et al., in samples from a laser capture microdissected prostate cancer dataset and a multi-cancer dataset. All signatures and sub-signatures generated from Varambally et al. were tested for enrichment in individual samples from two additional datasets – a laser-capture microdissected prostate cancer dataset generated by Tomlins et al., and a multi-cancer dataset generated by Ramaswamy et al., Column A and B are collectively signature identifiers indicating the threshold for coexpression (column A) and the signature/module number (column B). Column C indicates the overall direction of the fold-change indicating overexpression ('induced') or downregulation ('repressed'). Only signatures achieving a fold-change of at least two-fold are listed. Columns C through to CX represent individual samples from the Tomlins study. The data are listed sample-by-sample in order of clinical progression beginning with prostatic intraepithelial neoplasia ('PIN') cases and progressing to localised prostate cancer ('Primary Carcinoma'), followed by metastatic hormone naïve cases and finally metastatic hormone refractory samples. The GSM accession code provided for each sample is traceable through the NCBI GEO data repository from which the data were downloaded. Finally there are columns representing pairwise sample comparisons between metastatic hormone refractory and naïve samples. Fold-change information for the data for samples from Ramaswamy et al. is represented in column CY onwards (81k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s9.xlsx

Additional file 10: Table S8. Significance values and overall direction of the differential expression of gene signatures derived from Varambally et al., in samples from a laser capture microdissected prostate cancer dataset and a multi-cancer dataset. All gene signatures are classified with aggregate p-values, overall gene ontology assignments and a descriptor of overexpression ('induced') or underexpression ('repressed') normalised to benign specimens from GSE3325. Laser capture microdissected material is subgrouped as described in the original publication and in the methods section incorporating prostatic intraepithelial neoplasia (PIN), localised prostate cancer (PCA), hormone refractory (HR), metastatic hormone refractory (MET_HR) and metastatic hormone naïve disease (MET_HN). In addition gene signatures are defined as sub-signatures of larger signatures generated using lower stringencies/broader branching thresholds, the relationship between gene signatures at different thresholds is indicated in columns X-AE. The number of genes within each coexpression module is given in the column entitled 'moduleSize.varambally'. The numbers of corresponding mapped genes in the other data sets are indicated in columns Y and W. Gene ontology assignments for each signature are provided in column AF (40k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s10.xlsx

Additional file 11: Table S9. Gene composition and ontology assignments to statistically significant gene signatures. Sheet one lists the gene signatures and the genes that comprise them which are capable of clustering sample groups within the Tomlins and Ramaswamy datasets (Fischer's exact test < 0.05 p-value). Sheet two lists the GO terms associated with each significant discriminatory gene signature (<0.05 p-value). The discriminatory groups shown here mirror those depicted in the heatmap in Figure 1. Sheet three summarises the GO terms that are significant discriminators of important clinical sample groupings within the Tomlins and Ramaswamy datasets (550k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s11.xlsx

Additional file 12: Table S10. AUC values for individual probes for ERG, KLK3, AR and the gene signature in the test dataset from Grasso et al (1083k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s12.xlsx

Additional file 13: Table S11. AUC values for individual probes for ERG, KLK3, AR and the gene signature in the evaluation dataset from Grasso et al (44k)
http://www.biomedcentral.com/content/supplementary/s12920-014-0074-9-s13.xlsx